UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de la Telecomunicación

TRABAJO FIN DE GRADO

CLASIFICACIÓN DE GÉNEROS MUSICALES BASADA EN CONTENIDO

Ángel Pérez Lemonche
Tutor: Dr. Daniel Ramos Castro

JULIO 2014

CLASIFICACIÓN DE GÉNEROS MUSICALES BASADA EN CONTENIDO

Autor: Ángel Pérez Lemonche

Tutor: Dr. Daniel Ramos Castro

Trabajo de Fin de Grado

Grado en Ingeniería de Tecnologías y Servicios de la Telecomunicación

Área de Tratamiento de Voz y Señales

Departamento de Tecnología Electrónica y de las Telecomunicaciones

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Julio de 2014

RESUMEN

Hoy en día, la música en formato digital forma parte de nuestra vida cotidiana, desde la radio en Internet, hasta la compra *online* de canciones de nuestros artistas preferidos. En la red existen millones de repositorios no etiquetados de música, cuyo acceso sería mucho más rápido si existiese más información (metadatos) de esas canciones. También, los sistemas de recomendación relacionan las canciones que escuchamos y que más nos gustan con otras canciones que podrían ser de nuestro interés. Si conocemos el género musical de una canción partiendo de la información musical que nos proporciona ésta, tendríamos una primera aproximación a ambos problemas.

En este Trabajo de Fin de Grado se diseña y analiza la eficiencia de un clasificador de género musical basado en contenido, en el que se ha realizado un conjunto de pruebas cambiando diferentes parámetros para poder valorar cuáles de éstos son los que dan mejores resultados.

ABSTRACT

Nowadays, digital-format music takes part in our daily life, as Internet radio or online music stores that let us buy songs of our favorite artists from home. There are millions of non-tagged music databases in the Web and we could access them quicker if there were more information (metadata) about those songs. Also, recommendation systems relate the songs we listen to and we like best with another songs that could be interesting for us. If we know the musical genre of a song from the musical information of it, we will have a first approach to both problems.

In this Degree Project a content-based musical genre classifier has been designed and analyzed, a series of test have been done tuning different parameters to evaluate the ones that give the best performance.

PALABRAS CLAVE

Procesamiento de señal de audio, recuperación de información musical basada en contenido, clasificación de género musical, aprendizaje automático.

KEY WORDS

Audio signal processing, content-based Music Information Retrieval, musical genre classification, machine learning.

AGRADECIMIENTOS

Este Trabajo está dedicado enteramente a mi familia, principalmente a mis tres pilares, que me han apoyado para que siguiera adelante cuando parecía que no había luz al final del túnel.

Quiero expresar mi más sincero agradecimiento a mi tutor, Dr. D. Ramos, que me ha enseñado de forma muy amena la puerta que conduce a mi futuro. También quería dar las gracias a mis compañeros del laboratorio R. Landriz y F. Espinoza, por estar ahí cuando lo he necesitado y por la paciencia que han tenido conmigo. No puedo olvidar a mis profesores de estos últimos años, Dr. D. Torre, Dr. J. Bescós, Dr. J. Galbally, Dr. J. M.ª Martínez y Dr. J. González por despertarme el interés hacia el tratamiento digital de señales y el aprendizaje automático con su motivación y ejemplo.

Gracias también a mis compañeros de curso, en especial a Erik Velasco, Adrián Tomé, Paula Sánchez, Manuel Iglesias, Eduardo García, Álvaro Culebras, Alberto Palero, Alberto Mozo, Gabriel Álvarez, Ana Sotomayor, Sandra Gaytán, Diego Barrio, Víctor Hugo García, Carlos García, Sara Cerro, Patricia Marín, Marta Martínez, Sandra Jurado, Laura Usero, Jorge Sanjuan, Ana Huélamo, Adrián Cobos, Carlos Moreno, Ana Chevasco, Ignacio Ramos… Por haber hecho que estos cuatro años hayan sido especiales y menos duros día a día, gracias a vuestro buen humor, vuestra compañía y vuestras risas.

Por último, gracias a mis amigos de toda la vida, Jorge Arrieta, Alejandro Moreno, Rodrigo Mateos, Andrea Jurado, Andrés Hernández, Víctor Sanz, Marina Calleja, María Sacido, M.ª Paz Rubio, Ainhoa Morillas, Víctor Alonso, Pinar Sanz, Cristina García-Calvo y Angélica Ingelmo por estar ahí siempre.

ÍNDICE GENERAL

Tabla de contenido

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

GLOSARIO

- ***Beat***: se define como *beat* a la velocidad de una pieza musical, medido en pulsos por segundo. Es lo mismo que el *tempo* de una canción.
- ***Cepstrum:*** espacio resultado de aplicar la Transformada Inversa Discreta de Fourier (IDCT) al logaritmo del valor absoluto de la Transformada Discreta de Fourier (DFT) de una señal temporal.
- **CMS**: siglas de *Cepstral Mean Substraction*, traducido al español como *Resta de la media cepstral*. Técnica descrita en la Sección 4.3.2.
- **EM**: siglas de *Expectation Maximization*. Algoritmo iterativo utilizado para entrenar la mezcla de gaussianas en los modelos GMM-ML y GMM-UBM.
- ***Features***: traducido al español como *características*. En el contexto de este Trabajo, son valores distintivos de un tramo de una señal de audio que pretende ser similar en archivos de audio parecidos y muy diferente en señales de audio muy distintas.
- **GMM**: siglas de *Gaussian Mixtured Model*, traducido al español como *Modelo de mezclas de gaussianas*. Modelo estadístico descrito en la Sección 3.2.
- ***Ground truth*:** las etiquetas de *ground truth* son aquellas resultado de verificar manualmente la correspondencia de un dato a una categoría. En aprendizaje supervisado estas categorías son las del conjunto de entrenamiento y sirven para evaluar los resultados en la fase de test.
- ***K-means***: Algoritmo iterativo utilizado para inicializar los centros de las gaussianas de los modelos GMM-ML y GMM-UBM.
- ***LabROSA***: siglas de *Laboratory for the Recognition and Organization of Speech and Audio*. Proveen *software* para la extracción de características diseñado para señales de voz y audio. En este documento, cuando se hable de *LabROSA* se referirá exclusivamente al *software* para la extracción de características MFCC. http://labrosa.ee.columbia.edu/matlab/rastamat/
- **MAP:** siglas de *Maximum A Posteriori*. Método descrito en la Sección 3.2.
- ***Marsyas*:** entorno de *software* donde proveen la base de datos de géneros musicales utilizada en el Trabajo. En este documento, cuando se hable de *Marsyas* se referirá exclusivamente a la base de datos de géneros utilizada. http://marsyas.info/download/data_sets/
- **MFCC***:* siglas de *Mel-Frequency Cepstral Coefficients*. Método descrito en la Sección 3.1.

- **MIR:** siglas de *Music Information Retrieval*, traducido al español como *Recuperación de información musical*. Ciencia multidisciplinaria que se encarga de recuperar información a partir de los archivos de audio.
- **ML**: siglas de *Maximum Likelihood*, traducido al español como *Máxima verosimilitud*.
- **Netlab**: *toolbox* de MATLAB utilizado para el entrenamiento y el test de los modelos GMM-ML y GMM-UBM diseñados en este Trabajo.
  http://www.mathworks.com/matlabcentral/fileexchange/2654-netlab
- ***Pitch***: frecuencia fundamental de una ventana sonora.
- ***Score***: traducido al español como *puntuación*. Son los resultados de la fase de test de ambos modelos. Su cálculo se explica en la Sección 4.4.
- **UBM:** siglas de *Universal Background Model*, traducido al español como *Modelo universal*. Modelo estadístico descrito en la Sección 3.2.
- **ZCR:** siglas de *Zero Crossing Rate*, traducido al español como Tasa de cruces por cero. Técnica descrita en la Sección 4.3.2.

TRABAJO

SECCIÓN 1: INTRODUCCIÓN

En esta sección se explicará cual ha sido la motivación que ha llevado a realizar este Trabajo, así como los objetivos concretos, tanto principales como secundarios, que se han buscado alcanzar en el desarrollo de este Trabajo a lo largo de toda su duración y en fases concretas de éste. También se explica la estructura en la que está constituida la memoria, para hacer más fácil al lector guiarse por ella y comprenderla.

1.1 Motivación

Con la llegada de la llamada *Era Digital* y mediante el gran impulso que ha dado Internet de banda ancha al acceso de dichos contenidos digitales, los usuarios, nosotros, hemos cambiado la forma de adquirir la información que se encuentra en la Red. Poco a poco, los contenidos más demandados son aquellos que tienen contenido multimedia, tales como las imágenes, los vídeos, y cómo no, el audio.

Los métodos automáticos de recomendación son potentes herramientas que sugieren al usuario acceder a contenidos similares a aquel que actualmente se está reproduciendo. Podemos ver en grandes plataformas tales como *YouTube* y *Spotify* el gran éxito de estos sistemas de recomendación.

La clasificación de género musical basada en contenido es uno de estos métodos automáticos que consiguen extraer la información musical, no de las etiquetas que pudiera tener la canción, sino haciendo uso de la información inherente a la propia señal de audio. La clasificación de género es un mundo dentro del universo de la Recuperación de Información Musical (*Music Information Retirval*, MIR) [3] [4].

Debido a la ingente cantidad de audio (cuya duración ya se mide en años o siglos) no etiquetado que existe en bases de datos privadas y públicas, como las existentes en Internet, y el alto precio que cuesta la etiquetación de forma manual, se suele recurrir a métodos automáticos de etiquetado para facilitar el acceso y la velocidad de éste a la información. También, estos métodos cobran su importancia a la hora de funcionar como interfaz entre la realidad y el mundo a nivel de *bit*, ya que utiliza información acústica, traducida a digital, más sencilla de procesar, para obtener una información de más alto nivel: el género musical al que pertenece una canción.

Por último, cabe destacar la importancia de la clasificación de género no sólo en el marco del *Big Data*, como ya se ha mencionado anteriormente, sino en el ámbito de la musicología. Conforme los métodos digitales se desarrollan permiten avances en muchas ramas del conocimiento; y en el ámbito de la Música tiene utilidad, no sólo en la creación de nuevos sonidos, sino también en la clasificación de obras y canciones, ámbito en el que hay áridos debates entre musicólogos de todo el mundo.

1.2 Objetivos

El Trabajo se ha llevado a cabo con el fin de conseguir los siguientes objetivos.

- Diseñar un clasificador de género musical con una elevada eficiencia de clasificación en un tiempo razonable.
- Hacer uso de los métodos de reconocimiento de locutor con el fin de aplicarlos al reconocimiento de género musical.
- Aumentar la eficiencia de clasificación de un primer modelo haciendo uso de métodos que utilicen información tímbrica.
- Aumentar la eficiencia de clasificación haciendo uso de métodos que utilicen información no cepstral.
- Estar iniciado en el mundo del aprendizaje automático mediante la comprensión en profundidad de los métodos y técnicas utilizadas, así como los fundamentos matemáticos de los mismos y mediante la realización de un Trabajo que ha constado de tres etapas principales: estudio previo, realización y evaluación.

1.3 Estructura de la memoria

Esta memoria está redactada de acuerdo a la estructura establecida en la Normativa del Trabajo de Fin de Grado de la Escuela Politécnica Superior de la Universidad Autónoma de Madrid, aprobada en Junta de Centro el 30 de abril del año 2013, y está dividida en ocho partes: *Resumen*, Abstract, *Palabras clave* / Key words, *Agradecimientos*, *Índices*, *Glosario*, *Trabajo* y *Bibliografía*. Dentro del apartado denominado *Trabajo* se desarrollan seis secciones:

Sección 1: Introducción. En esta primera sección se ha expuesto tanto la motivación como los objetivos del Trabajo desarrollado, así como también este apartado en que se explica brevemente cómo está estructurada toda la memoria.

Sección 2: Estado del arte. En esta sección se cuenta la situación en la que se encuentra la clasificación de género musical basada en contenido dentro del marco de la recuperación de información musical (*Music Information Retrieval,* MIR).

Sección 3: Métodos utilizados. En esta sección se exponen las técnicas y métodos utilizados para el desarrollo del sistema clasificador, haciendo alusión a las referencias bibliográficas fundamentales donde aparecen publicadas dichas tecnologías.

Sección 4: Diseño y desarrollo. En esta sección se detallan las partes que forman el clasificador de género musical basado en contenido, así como los parámetros escogidos para su funcionamiento. También se explican qué funciones han sido utilizadas de los paquetes de *software* abierto de *LabROSA* y *Netlab*[1], y el funcionamiento de todas las herramientas implementadas.

Sección 5: Pruebas y resultados. En esta sección se analizan las pruebas realizadas con las variaciones que se consideraron oportunas para conseguir mejorar el sistema de clasificación, así como los resultados de cada prueba y un análisis de resultados entre las pruebas.

Sección 6: Conclusiones y trabajo futuro. En esta última sección se valoran en conjunto los resultados obtenidos y se da una explicación a los mismos. También se exponen diferentes líneas de trabajo en las que se podría actuar para mejorar los resultados del clasificador diseñado y presentado en esta memoria.

[1] *LabROSA* y *Netlab* son dos paquetes *software* libre referenciados en la Sección 4: Diseño y desarrollo.

SECCIÓN 2: ESTADO DEL ARTE

La música es una manifestación artística muy popular en todas las culturas de la humanidad. En el siglo XXI, donde podemos entender la cultura como un conjunto de conocimientos globalizados, debido al extendido uso de Internet y las facilidades que nos proporciona la *World Wide Web* para la compartición de datos, la música sigue teniendo una influencia importante, siendo uno de los tipos de información *online* más populares tanto en portales de descargas como *ITunes* o en reproducción en *streaming* como *Spotify*.

Debido a la existencia de colecciones con más de decenas de millones de canciones [3] y a las facilidades que nos ofrecen los sistemas automáticos de búsqueda, de recuperación de información, de catalogación de archivos de audio o de recomendación por contenido, muchos investigadores de todo el mundo se plantean el reto de desarrollar estas tecnologías para hacer frente al problema del *Big Data* en el entorno de la información de audio o musical.

La recuperación de información musical (*Music Information Retrieval*, MIR en adelante) recoge un conjunto de estrategias para poder acceder a las colecciones de música, tanto nuevas como históricas, que necesitan ser desarrolladas con el fin de mantener unas expectativas en función de la aplicación que quiera darse, como por ejemplo el tiempo de búsqueda, precisión en la búsqueda, *ranking* de búsquedas similares, ... Los destinatarios habituales que utilizan sistemas MIR son las industrias musicales, musicólogos, músicos, profesores, abogados encargados del *copyright*, productores de música y usuarios finales que quieren encontrar música con fines particulares.

Los estudios e investigaciones en el campo de MIR han juntado a expertos de diversas disciplinas tales como musicólogos, expertos en percepción, psicólogos del campo de la cognición, ingenieros e informáticos, que juntos han trabajado para encontrar soluciones a este problema usando métodos basados en contenido.

Pese a que los métodos más comunes de búsqueda de música se realizan a través de metadatos textuales, los métodos basados en contenido son muy potentes, ya que identifican lo que un usuario quiere encontrar incluso si éste no sabe específicamente lo que está buscando. Además de esto, las descripciones mediante metadatos son muy costosas, ya que requieren que un experto genere etiquetas de verdad asentada o etiquetas de *ground truth*, y este proceso puede llevar unos 20 ó 30

minutos por canción. Otra alternativa también utilizada para la generación de metadatos es la del uso de sistemas públicos de votación usando redes sociales, pero aun así existe el problema de la no uniformidad de intereses de los usuarios y de que estos descriptores pueden representar opiniones, ya que no están supervisados por expertos.

Por todo ello, existe una tendencia hacia el desarrollo de métodos automáticos para la generación de estos metadatos, además de la tendencia a la realización de búsquedas a más alto nivel utilizando métodos de búsqueda mediante consultas no textuales.

Existen diferentes niveles de especificidad a la hora de buscar o encontrar información de una canción. En [3] se definen tres niveles: alto (*H*) medio (*M*) y bajo (*L*) en función de la exactitud requerida para resolver el problema. La clasificación de género musical es uno de los casos de uso de las herramientas MIR, y se considera de baja especificidad, como muestra la **Tabla I**, ya que para que el sistema funcione correctamente no se necesita que el resultado sea exacto, debido a que puede bastar para una aplicación concreta que el resultado sea aproximado. Además de esto, cuando se realiza una búsqueda, para una consulta se pueden obtener más de un resultado.

Use Case	Speci-ficity	Description
Music Identification	H	Identify a compact disk, provide metadata about an unknown track, mobile music information retrieval: e.g. *shazam.com*
Plagiarism detection	H	Identify mis-attribution of musical performances, mis-appropriation of music intellectual property.
Copyright monitoring	H	Monitor music broadcast for copyright infringement or royalty collection
Versions	H/M	Remixes, live vs. studio recordings, cover songs. Used for database normalization and near-duplicate results elimination
Melody	H/M	Find works containing a melodic fragment
Identical Work / Title	M	Retrieve performances of same opus number or song title
Performer	M	Find music by a specific artist
Sounds like	M	Find music that sounds like a given recording
Performance Alignment	M	Mapping one performance onto another independent of tempo and repetition structure
Composer	M	Find works by one composer
Recommend-ation	M/L	Find music that matches the user's personal profile
Mood	L	Find music using emotional concepts: *Joy, Energetic, Melancholy, Relaxing*
Style / Genre	L	Find music that belongs to a generic category: *Jazz, Funk, Female Vocal*
Instrument(s)	L	Find works with same instrumentation
Music-Speech	L	Radio broadcast segmentation, Music archives cataloguing

Tabla I, Especificidad en MIR [3]

La baja especificidad de las tareas de clasificación de género también va sujeta a la propia naturaleza de los géneros musicales —no todas las canciones pueden definirse en un género musical concreto: algunas canciones pueden pertenecer a varios géneros de forma simultánea, otras son generatrices de nuevos géneros que se especifican con el paso de los años—. También es importante mencionar la complejidad de la clasificación de género no sólo entre ellos, sino también dentro de uno mismo, con canciones que son de diferentes estilos musicales —no se puede considerar que es lo mismo un *Rock and Roll* de Elvis Presley que el *Rock Psicodélico* que hicieron Jefferson Airplane o el *Rock Alternativo* de Nirvana—.

Lo que sí que se puede afirmar de manera comúnmente consensuada es que, desde el punto de vista cognitivo y perceptual, existe una correspondencia entre el género de una canción y la textura musical de ésta, también llamada timbre. El timbre es aquella característica que distingue dos sonidos del mismo tono y misma intensidad y que, a nivel espectral, viene determinado por la forma de los armónicos y, sobre todo, por la relación entre ellos, apreciable en la envolvente espectral a corto plazo. Los seres humanos hacemos relaciones entre canciones según su timbre, es decir, según cómo suenan éstas [2]. Para parametrizar el timbre se suele hacer uso de extracciones de características de bajo nivel que modelan la envolvente espectral [3], siendo los *Mel-Frecuency Cepstrum Coefficients* (MFCC) los más extendidos en el estado del arte [2], y por ello, este Trabajo hace uso de ellos. A parte de la información tímbrica, también otro tipo de características como el *beat*, la información de *tempo*, la segmentación de voz cantada o los metadatos ayudan a mejorar la eficiencia de clasificación.

En [2], un *paper* muy referenciado en este Trabajo y de gran importancia en el área MIR, se contribuye al estado del arte exponiendo que a partir de una determinada complejidad, la eficiencia en la clasificación llega a un límite, un "techo de cristal", como muestra la **Figura 1**, esto es, cuesta más trabajo extraer ciertas características, o el entrenamiento de un modelo más complejo, para la mejora parcial o global que pueda suponer su uso. Por otro lado, en relación a los modelos, este *paper* concluye con la siguiente observación:

> *"Cognitive evidence show that human subjects tend not to assess similarity by testing the significance of the hypothesis 'this sound like X', but rather by comparing two competing models 'this sounds like X' and 'this doesn't sounds like X'"* [2]

Por ello, en este Trabajo también se ha hecho uso de un modelo universal (*Universal Background Model*, UBM [11]), ampliamente utilizado en el área de reconocimiento de locutor, que acentúa las diferencias entre canciones que se parecen a un determinado género y otras que no se parecen a otro determinado género.

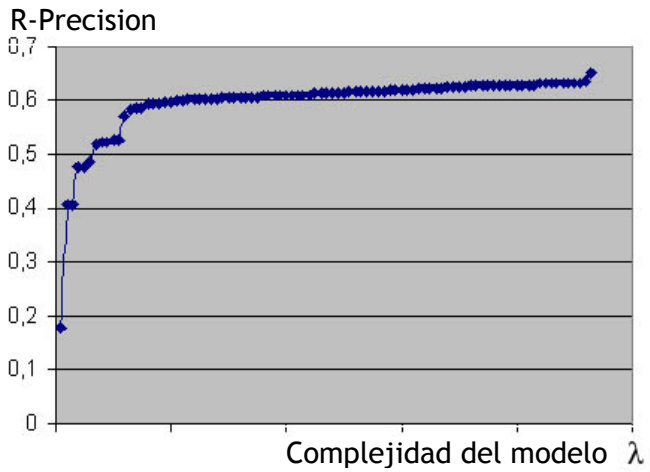


Figura 1, Techo de cristal [2]

Para concluir con el estado del arte, es indispensable dedicar unas palabras a MIREX (*Music Information Retrieval Evaluation eXchange*), una evaluación anual de los algoritmos del área de MIR, que tiene lugar junto con las conferencias ISMIR (*International Symposium on Music Information Retrieval,* [5]), donde los participantes pueden proponer los temas que se quieren tratar y donde los organizadores de MIREX ejecutan los algoritmos en ordenadores locales con colecciones musicales que no distribuyen.

En cuanto a clasificación de género, los algoritmos más comúnmente utilizados[2] son los que se basan en MFCCs y GMMs, utilizando también caracterizaciones de texturas tímbricas basadas en la *Short-Time Fourier Transform*, tales como centroides espectrales, dispersiones espectrales, brillo espectral, entropía espectral, etcétera. En los últimos años también se ha valorado utilizar *Support Vector Machines* (SVM) con *kernels* de *Gaussian Radial Basis Functions* (RBF) como modelo estadístico.

En MIREX 2005 [13] tenemos una evaluación de un algoritmo con una tasa de aciertos general del 75.29%. Este Trabajo ha sido enfocado para conseguir un resultado similar, haciendo uso de un conjunto de herramientas que se han evaluado en numerosos estudios e investigaciones y que constituyen el estado del arte.

[2] Información extraída de la página oficial de MIREX, evaluaciones del año 2013: http://www.music-ir.org/mirex/wiki/MIREX_HOME

SECCIÓN 3: MÉTODOS UTILIZADOS

En esta sección se explican los métodos matemáticos fundamentales utilizados a lo largo del desarrollo de este Trabajo. Está dividida en tres partes:

Extracción de características: se explica cómo se extraen los coeficientes MFCC de un tramo de señal de audio.

Modelado estadístico: se explica en qué consiste el modelo GMM, UBM y cómo se realiza la adaptación MAP.

Normalización de puntuaciones: se explica los dos tipos de normalizaciones utilizadas: la Normalización T y la Normalización Z.

3.1 Extracción de Características

A continuación se describe el procedimiento del cálculo de las características utilizadas en este Trabajo, que fundamentalmente son los *Mel-Frequency Cepstral Coefficients* (MFCCs). El resto de técnicas utilizadas como características se describen en la Sección 4.3.2 ya que se basan en los MFCCs para su cálculo. Los MFCCs, como el resto de las características seleccionadas, se extraen sobre una ventana temporal, descrita con mayor detalle en la Sección 4.3.1.

Mel-Frequency Cepstral Coefficients (MFCC)

Los MFCCs son coeficientes cepstrales derivados del análisis sobre la escala Mel y su uso está ampliamente extendido en el tratamiento digital de voz y audio [2].

Existen dos formas de calcular los MFCCs, de las cuales sólo se explica aquí el segundo método, que es el más comúnmente utilizado:

- Aplicando la Transformada Discreta de Fourier (DFT) sobre una señal sobremuestreada, con el fin de tomar los coeficientes más cercanos a las frecuencias centrales de las bandas Mel (conversión escala de frecuencia f a escala Mel m en la **Ecuación 3.1**).

$$m = 1127{,}01048 \cdot \log_e\left(1 + \frac{f}{700}\right) \qquad (3.1)$$

- Basándose en la energía total logarítmica de cada banda crítica.

Se define $Y(i)$, energía total logarítmica de la banda i, B_i, que se calcula de la siguiente manera:

$$Y(i) = \sum_{k \in B_i} \log|S(k)| \cdot H_i\left(\frac{2\pi k}{N}\right) \qquad (3.2)$$

Donde $S(k)$ es el valor de la DFT de una ventana de la señal de audio, N es el número de puntos para el cálculo de la DFT y $H_i\left(\frac{2\pi k}{N}\right)$ nos indica la frecuencia central normalizada de la banda i de cada uno de los filtros triangulares en escala Mel, que se representa en la **Figura 2**.

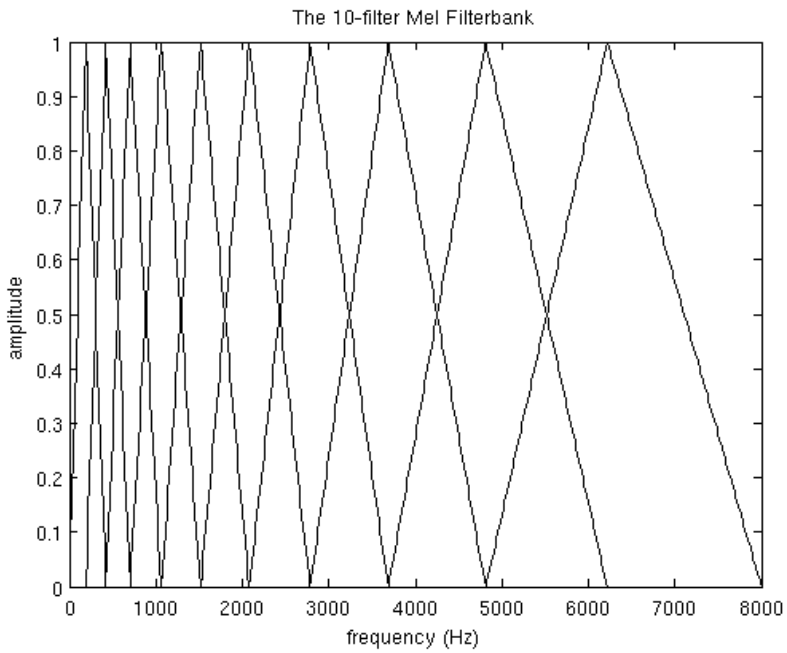


Figura 2, Banco de filtros en escala Mel.
***http://practicalcryptography.com/*.**

Posteriormente, se define la secuencia de energía logarítmica espectral de la siguiente manera:

$$\widetilde{Y}(k) = \begin{cases} Y(i), & k = k_i \\ 0, & \text{resto} \end{cases} \qquad (3.3)$$

Debido a que la secuencia $\widetilde{Y}(k)$ es simétrica respecto a $N/2$, podemos reemplazar la exponencial para transformar la secuencia al dominio cepstral por la función coseno. Por otro lado, ya que existen tantos valores distintos de cero en el sumatorio como número de filtros de bandas Mel utilizados N_B, la Transformada Inversa Discreta de Fourier (IDFT) de la secuencia $\widetilde{Y}(k)$, y, por tanto, la expresión de los coeficientes MFCC, resulta de la siguiente manera:

$$c(n) = \frac{2}{N} \cdot \sum_{i=1,2,\dots,N_B} \widetilde{Y}(k_i) \cdot \cos(n k_i \frac{2\pi}{N}) \qquad (3.4)$$

Una propiedad interesante del dominio cepstral es que los coeficientes cepstrales tienden a presentar ortogonalidad, lo que puede simplificar en ocasiones el modelado estadístico.

3.2 Modelado estadístico

Para el modelado estadístico se utilizan en este trabajo dos métodos para modelar los datos: *Gaussian Mixture Model* (GMM) y *Universal Background Model* (UBM). Para el segundo, se adaptan los modelos mediante la adaptación *Maximum A Posteriori* (MAP).

Gaussian Mixture Model (GMM)

El modelo de mezcla de gaussianas [11] es una manera de representar la función de densidad de probabilidad de una función como la suma de densidades de componentes gaussianas.

Un GMM es la suma ponderada de las densidades de M componentes gaussianas:

$$p(v|\lambda) = \sum_{i=1}^{M} w_i \cdot g(v|\mu_i, \Sigma_i) \qquad (3.5)$$

Donde v es una observación o, en este caso, un vector de características N-dimensional de valores continuos, w_i, con $i = 1, \dots, M$ son los pesos de la mezcla, que satisfacen que $\sum_{i=1}^{M} w_i = 1$, y $g(v|\mu_i, \Sigma_i)$ es la densidad de cada componente gaussiana, definida como:

$$g(v|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{N/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \qquad (3.6)$$

Donde μ_i es la vector de medias y Σ_i la matriz de covarianzas de cada componente gaussiana.

El modelo, λ, se caracteriza por el conjunto de parámetros necesarios para determinar su función de densidad de probabilidad: $\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, \dots, M$.

La **Figura 3**, muestra un GMM de $M = 4$ componentes con $v \in \mathbb{R}$. Cada componente gaussiana se representa con una línea roja, mientras que la función de densidad de probabilidad del GMM completo corresponde con la línea azul.

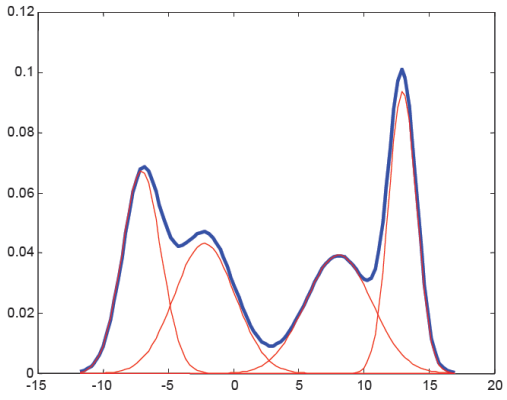


Figura 3, GMM M=4 sobre 1D. ***Material docente de la asignatura Tratamiento Digital de Voz (EPS-UAM)***

Por otro lado, la **Figura 4**, muestra otro GMM de $M = 4$ componentes con $v \in \mathbb{R}^2$, y la **Figura 5** representa las curvas de nivel para un valor de $p(v|\lambda)$ fijo.

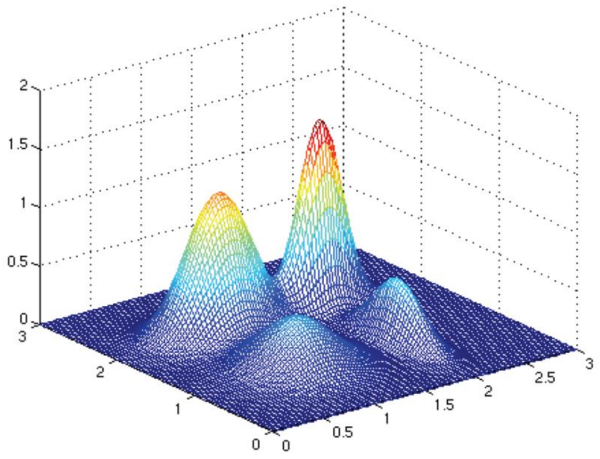


Figura 5, GMM M=4 sobre 2D. *Material docente de la asignatura Tratamiento Digital de Voz (EPS-UAM)*

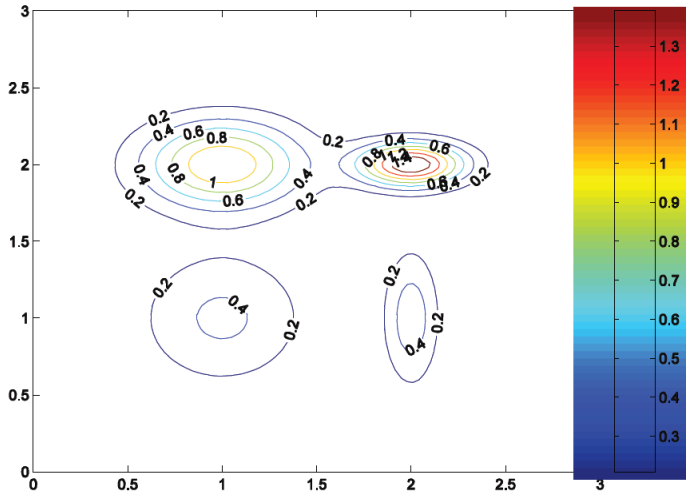


Figura 4, Curvas de nivel GMM M=4 sobre 2D. Material docente de la asignatura Tratamiento Digital de Voz (EPS-UAM)

Para un conjunto de vectores de características, se asume independencia entre vectores, obteniendo así la verosimilitud de que ese conjunto de vectores pertenezcan a un modelo:

$$p(V|\lambda) = \prod_{t=1}^{T} p(v_t|\lambda) \qquad (3.7)$$

Siendo V el conjunto de T vectores de características.

Para el entrenamiento de este modelo se utiliza el algoritmo *Expectation Maximization*[3].

[3] Para más información acerca del algoritmo *Expectation Maximization* consultar [6] o la siguiente referencia: DEMPSTER, A et al. (1977): *Maximum Likelihood from Inclomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society 39(1) (1997): 1-38.

Debido a que, por problemas de coste computacional, el entrenamiento de matrices de covarianzas es muy elevado, se tiende a usar matrices de covarianza diagonales, es decir, utilizando sólo los valores de la varianza de cada variable y poniendo a cero todas las covarianzas. De esta forma, utilizando un número de componentes gaussianas suficiente se puede compensar la falta de matrices completas, adaptándose, las gaussianas, a las funciones de densidad de probabilidad que se desea modelar, como se muestra en la **Figura 6**.

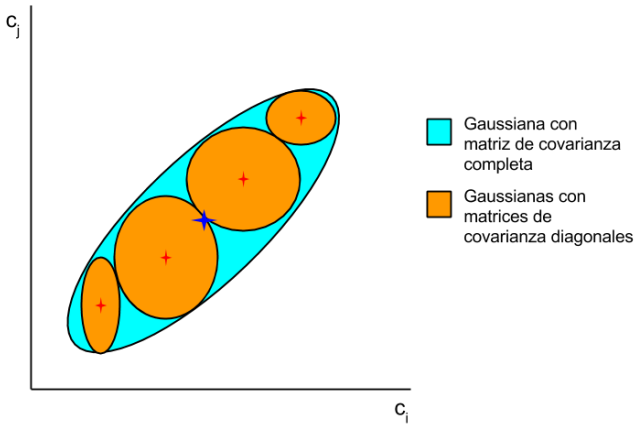


Figura 6, Curvas de nivel - Covarianzas de gaussianas

En la **Figura 7**, a la izquierda se muestra una gaussiana con matriz de covarianza completa y a la derecha, la misma función densidad de probabilidad modelada por cuatro componentes gaussianas con matrices de covarianza diagonales.

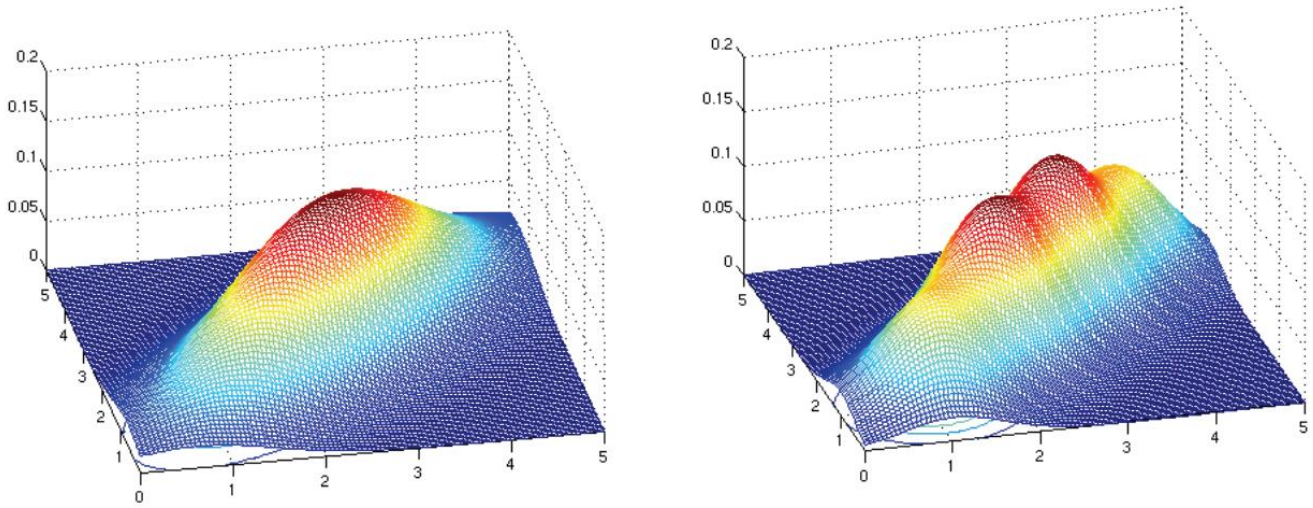


Figura 7, Representación en 3D - Covarianzas de gaussianas. ***Material docente de la asignatura Tratamiento Digital de Voz (EPS-UAM)***

Universal Background Model (UBM)

El modelo universal [11] es un modelo basado en GMM extendido en el uso de sistemas de reconocimiento biométrico, utilizado para representar de forma general un conjunto de datos sin tener en cuenta características independientes de éstos, es decir, en este caso, las características propias de un género musical en concreto. Por ello, aquí, el modelo universal también se denomina *modelo de música*, porque se modela utilizando todos los datos de entrenamiento, independientemente del género musical al que pertenezcan.

En la **Ecuación 3.8** se define la relación de verosimilitudes (*Likelihood-Ratio*), $\mathrm{LR}(\mathrm{V})$, dada una observación V —una canción de la base de datos de test, en el contexto de este Trabajo—, y dado un modelo de género λ_j y un modelo de género universal λ_U.

$$\mathrm{LR}(\mathrm{V}) = \frac{p(\mathrm{V}|\lambda_j)}{p(\mathrm{V}|\lambda_U)} \qquad (3.8)$$

Este valor será tanto mayor cuanto más se parezca una canción V a un modelo de género λ_j. De este modo, la relación de verosimilitudes tiende a puntuar por las diferencias con respecto al *modelo de música*, lo cual normaliza las puntuaciones del sistema para géneros diferentes.

Para adaptar los modelos de género a partir del modelo universal se suele emplear la técnica MAP (*Maximum A Posteriori*) [11]. Como muestra la **Figura 8**, a partir del UBM, utilizando los datos de entrenamiento de cada género musical se ajustan las componentes gaussianas a estos nuevos datos.

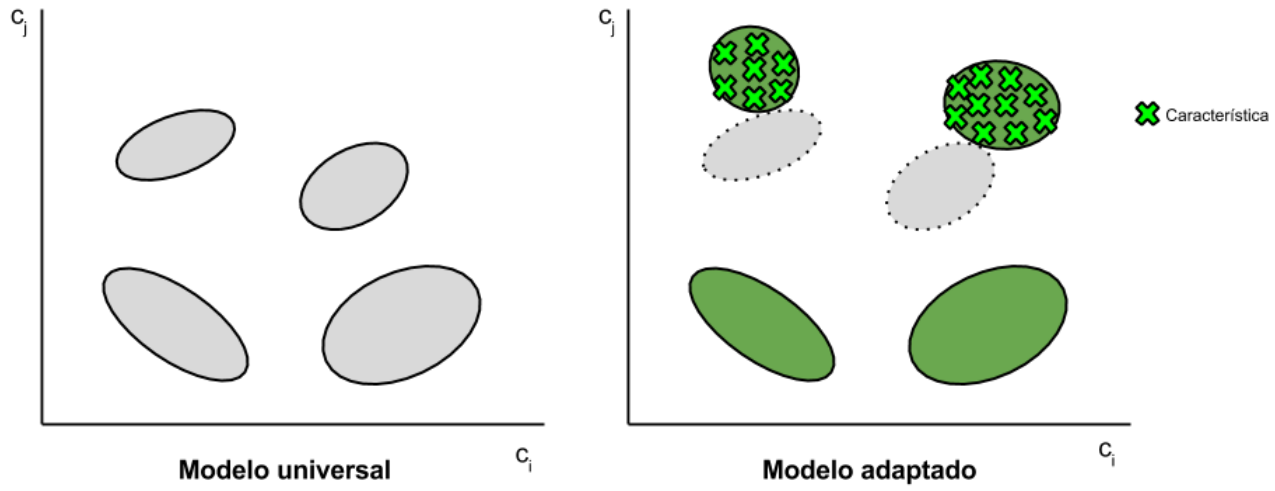


Figura 8, Adaptación MAP

MAP adapta las componentes gaussianas, es decir, pesos, medias y/o matrices de covarianza.

1. El algoritmo, en un primer paso, estima los valores de los estadísticos suficientes (*sufficient statistics)*, es decir, aquellos estadísticos básicos necesarios para calcular los parámetros deseados de los datos de entrenamiento para cada mezcla en el modelo *a priori*, es decir, el modelo previo a la adaptación, el modelo universal.
2. Después, en un segundo paso, combina los estadísticos suficientes "nuevos" —calculados en el paso anterior— con los "viejos" —de la mezcla *a priori*—. La combinación se hace de forma que las mezclas con un alto número de datos nuevos otorguen más relevancia a los estadísticos suficientes nuevos, mientras que las mezclas con un bajo número de datos nuevos tengan más en consideración los estadísticos suficientes antiguos a la hora de la estimación final de parámetros.

Para el paso 1 del algoritmo MAP, se calcula la probabilidad *a posteriori* de la componente gaussiana *i-ésima* de la siguiente manera:

$$\Pr(i|v_t, \lambda_{priori}) = \frac{w_i \cdot g(v_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k \cdot g(v_t|\mu_k, \Sigma_k)} \quad (3.9)$$

Donde aquí λ_{priori} es el modelo *a priori*, es decir, el modelo universal.

Con la probabilidad *a posteriori*, se calculan los estadísticos suficientes para los parámetros de pesos, medias y varianzas, haciendo para ello uso de las **Ecuaciones 3.10**, **3.11** y **3.12** respectivamente[4]:

$$n_i = \sum_{t=1}^{T} \Pr(i|v_t, \lambda_{priori}) \quad (3.10)$$

$$E_i(v) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|v_t, \lambda_{priori})\, v_t \quad (3.11)$$

$$E_i(v^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|v_t, \lambda_{priori})\, v_t^2 \quad (3.12)$$

[4] v^2 es la abreviatura de $\mathrm{diag}(vv')$

Para el paso 2 del algoritmo MAP, estos estadísticos suficientes son utilizados para actualizar los estadísticos suficientes *a priori* de la mezcla i, con el fin crear los parámetros adaptados para la mezcla i, utilizando las **Ecuaciones 3.1**, **3.14** y **3.15** para adaptar respectivamente pesos, medias y varianzas:

$$\hat{w}_i = [\alpha_i^w n_i/T + (1 - \alpha_i^w) w_i]\gamma \qquad (3.13)$$

$$\hat{\mu}_i = \alpha_i^m E_i(v) + (1 - \alpha_i^m)\mu_i \qquad (3.14)$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(v^2) + (1 - \alpha_i^v)\left(\sigma_i^2 + \mu_i^2\right) - \hat{\mu}_i^2 \qquad (3.15)$$

Donde γ es un factor de escala necesario para que la suma de los pesos sea unitaria. El coeficiente dependiente de los datos $\alpha_i^\rho, \rho\epsilon\{w, m, v\}$ que se utiliza en las tres ecuaciones anteriores se calcula de la siguiente manera:

$$\alpha_i^\rho = \frac{n_i}{n_i + r^\rho} \qquad (3.16)$$

Donde r^ρ es un factor de relevancia fijado. Normalmente se utiliza $r^\rho = r$ igual para el cálculo de los pesos, medias y varianzas. El factor de relevancia utilizado ha sido $r^\rho = r = 16$, extendido en el campo de reconocimiento de locutor [12].

De esta forma, si la probabilidad *a posteriori* es alta, $\alpha_i^\rho \to 1$, es decir, los estadísticos suficientes se adaptan. En cambio, si ésta es pequeña, $\alpha_i^\rho \to 0$ y los estadísticos suficientes permanecen como estaban en el modelo *a priori*.

3.3 Normalización de puntuaciones

El uso de la normalización de puntuaciones (*Score Normalization*) es una técnica empleada de forma habitual a la hora de trabajar con puntuaciones, ya que algunos experimentos muestran mejoras significativas utilizando estas técnicas [1], ya que corrige un problema de alineamiento entre diferentes modelos o ficheros de test.

Normalización T (T-Norm)

La Normalización T, *Test Normalization* o T-Norm es una técnica de normalización de puntuaciones basada en la estimación de la media y la varianza. Dado un *set* de puntuaciones, calculadas a partir de la verosimilitud de distintos modelos cuando se comparan con un segmento de audio de prueba, se calculan la media y la desviación típica de las puntuaciones de ese segmento con el resto de modelos, es decir, del resto de las puntuaciones del *set*, y luego se normaliza según la **Ecuación 3.17**.

$$\mathrm{score}_{\mathrm{Norm}} = \frac{\mathrm{score} - \mu_{\mathrm{I}}}{\sigma_{\mathrm{I}}} \qquad (3.17)$$

Donde score es la puntuación obtenida previa normalización, μ_{I} es la media de impostor y σ_{I} la desviación típica de impostor dado un modelo.

Normalización Z (Z-Norm)

La Normalización Z, *Zero Normalization* o Z-Norm es otra técnica de normalización de puntuaciones basada en la estimación de la media y la varianza. La diferencia que existe con la Normalización T es que los parámetros de normalización se hallan durante la fase de entrenamiento (ventaja si se quiere hacer una aplicación a tiempo real). En este Trabajo se ha tomado un 33% de la base de datos de entrenamiento (1 de cada 3 canciones) para el cálculo de estos parámetros, más concretamente, tomando la primera canción del primer género, la cuarta, la séptima, etcétera, hasta el final de la base de datos de cada género y para todos los géneros. Con estos ficheros se generan las puntuaciones de impostor, enfrentando dichos segmentos de audio a los modelos de género que no son de su género. Una vez calculados los parámetros necesarios para normalizar (media y varianza), se aplica la **Ecuación 3.17** a todas las puntuaciones de test que se generen en el sistema.

SECCIÓN 4: DISEÑO Y DESARROLLO

En esta sección se expone de forma descendente, es decir, de mayor abstracción a mayor detalle y de manera más concreta el funcionamiento del sistema diseñado para clasificar los géneros musicales basándonos en la información de la señal de audio.

En primer lugar, se presenta el clasificador, su estructura y las partes en las que está dividido. Posteriormente se explican detalladamente las cuatro partes del clasificador: la base de datos, el extractor de características, los modelos utilizados, y por último se analiza también cómo está constituido el evaluador de resultados, es decir, los criterios utilizados para valorar la bondad de un clasificador resultante en una prueba con respecto a otras pruebas realizadas.

Este clasificador ha sido desarrollado íntegramente en código MATLAB.

4.1 Descripción del sistema

El clasificador es el sistema principal utilizado para la clasificación automática de audio en función del género. Su funcionamiento se basa en tomar como entrada canciones de la base de datos y, por una parte, con un conjunto de canciones denominadas *de entrenamiento*, se entrenan varios modelos de género musical —en este sistema se trabaja con 10 géneros diferentes— y por otra parte, con otro conjunto diferente de canciones denominadas *de test* se evalúa la capacidad de clasificación del sistema. En el caso de este sistema se mide el porcentaje de canciones que han sido clasificadas correctamente.

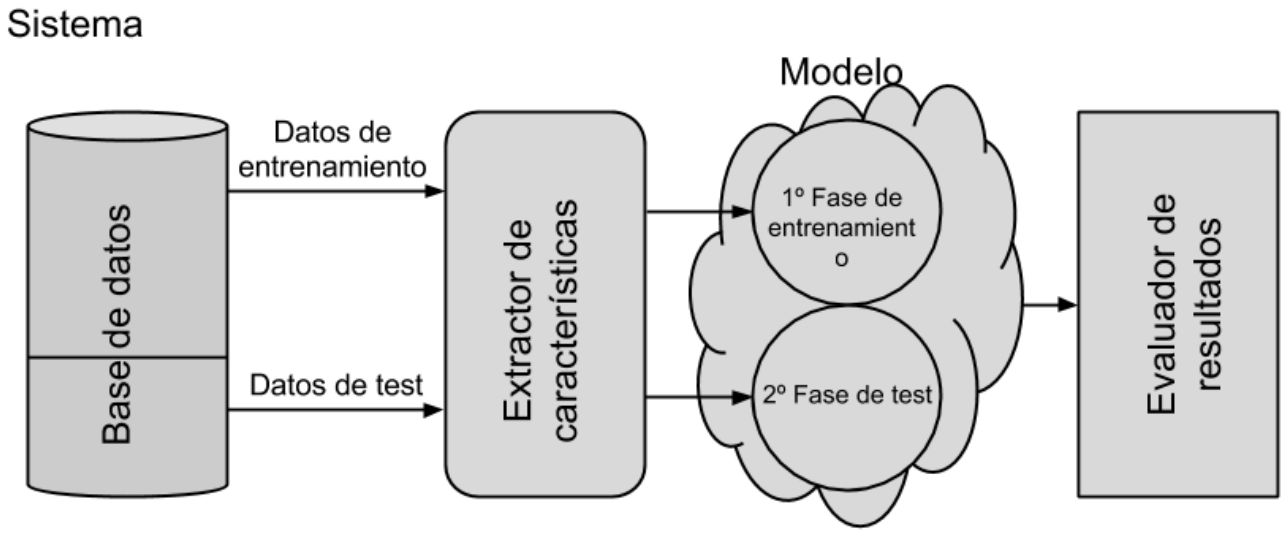


Figura 9, Estructura general del sistema

El clasificador consta de cuatro bloques principales, como se representa en la **Figura 9**:

La base de datos. Es el conjunto de datos utilizado. En este sistema está compuesta por un conjunto de 1.000 canciones en formato de forma de onda, 100 canciones por género musical (10 géneros musicales). Las canciones de la base de datos tienen dos funcionalidades principales: por una parte sirven para entrenar los modelos de género, y por otra parte, para evaluar la precisión de los diferentes modelos y las características extraídas de las canciones para entrenar los modelos.

El extractor de características. Las canciones de la base de datos se hacen pasar por un extractor de características (*features*) con el fin de utilizar una versión resumida y potente de las canciones, y así poder clasificar de forma más rápida y eficiente que si se utilizase la información de la forma de onda directamente, ya que también permite eliminar ruido si estas características son robustas —disminuye el error de clasificación—.

El modelo. Es el conjunto de métodos y técnicas utilizadas para conseguir representar los géneros musicales. Los modelos utilizados en este sistema son GMM (*Gaussian Mixture Model*) usando uso del criterio de máxima verosimilitud (*Maximum Likelihood,* ML) y el de UBM (*Universal Background Model*) utilizando adaptación MAP (*Maximum A Posteriori*).

El evaluador de resultados. En este bloque se escoge un criterio de evaluación de resultados con el fin de discriminar las pruebas realizadas en función de los resultados y tener un criterio medible para evaluar la bondad de los modelos.

A continuación se expone de forma más profunda y detallada cómo están constituidos los diferentes bloques, explicando también las características propias de la implementación utilizada.

4.2 La base de datos

La elección de la base de datos a utilizar por el sistema se ha hecho con diversos objetivos: en primer lugar, que los datos, es decir, la cantidad de canciones, sea lo suficientemente grande como para que los resultados puedan ser lo más genéricos posibles y no errar a la hora de diseñar un sistema que pueda dar resultados muy variables en distintos escenarios. En segundo lugar, se ha buscado una base de datos con diversidad de géneros musicales, para poder representar un conjunto amplio de géneros. También se ha escogido una base de datos donde los ficheros de audio se encuentren en formato de forma de onda, con el fin de extraer parámetros no solamente cepstrales, que son los más comúnmente utilizados en este tipo de clasificadores, sino tener abierta la oportunidad de extraer otro tipo de características, como la tasa de cruces por cero, y evaluar también el espectrograma de la señal.

La base de datos utilizada ha sido la base de datos pública *Marsyas* [5](*Musical Genre Classification of audio Signals*), en concreto la denominada *GTZAN Genre Collection* donde según afirma el autor, George Tzanetakis, fue constituida durante los años 2000 y 2001, utilizando archivos procedentes de *CDs*, de radio y grabaciones con micrófono, con el objetivo de representar diferentes condiciones de grabación.

Esta base de datos consta de 1.000 archivos de audio de 30 segundos de duración, divididos de forma equitativa en 10 géneros (100 canciones por género). Las pistas de audio están muestreadas a 22.050 Hz, en un único canal mono de 16 bits en formato *.au*. Los géneros musicales de esta base de datos son: *blues*, *classical*, *country*, *disco*, *hip hop*, *jazz*, *metal*, *pop*, *reggae* y *rock*.

Con vistas a realizar una clasificación basada en el entrenamiento de un modelo para posteriormente someter dicho modelo a una fase de test y así comparar resultados, se ha elegido tomar el 60% de la base de datos como datos de entrenamiento, y el 40% restante como datos de test. El criterio utilizado ha sido tomar las primeras 60 canciones de cada género como datos de test y las otras 40 como datos de entrenamiento, con el objetivo de tener una cantidad de datos suficiente, equitativa y característica de cada género.

[5] Marsyas es la base de datos utilizada para este Trabajo. Se puede acceder a ella en http://marsyas.info/download/data_sets/ .

4.3 Las características

Cuando los seres humanos afirmamos que una canción pertenece a uno u otro género musical cabe pensar que dicha canción que estamos escuchando tiene unas características comunes con aquellas que sabemos que pertenecen a ese género. Así pues, si escuchamos una canción donde el vocalista no canta, sino recita y existe una base musical de fondo constante donde predomina el sonido percusivo, podríamos clasificarla sin miedo a equivocarnos en el género *hip hop*. Por otro lado, si son los solos de guitarra ligeramente distorsionada una de las características más llamativas de la canción, podríamos decir que se trata de una canción de *rock*.

Es razonable pensar que estas características que el ser humano extrae de las canciones de manera cognitiva están estrechamente ligadas con el timbre de la canción, es decir, con las diferentes resonancias existen en la señal musical. Estas resonancias están directamente relacionadas con la respuesta en frecuencia de la señal de audio.

Como se puede observar en la **Figura 10**, utilizando el ejemplo citado anteriormente entre el *hip hop* y el *rock*, se aprecia claramente que las canciones del mismo género tienen espectrogramas similares, pero cuando los comparamos entre géneros distintos apreciamos diferencias. En este caso, las canciones de *hip hop* tienen evidentes muestras de componentes periódicas de tipo ruidoso —líneas verticales— que podríamos asociarlas con los golpes de percusión. En contraposición, las canciones de *rock* tienen más energía, no sólo de forma general, también se encuentra repartida a lo largo de todo el espectro debido al uso de diferentes instrumentos musicales, tanto de bajas como de medias y altas frecuencias, como el bajo y la guitarra respectivamente.

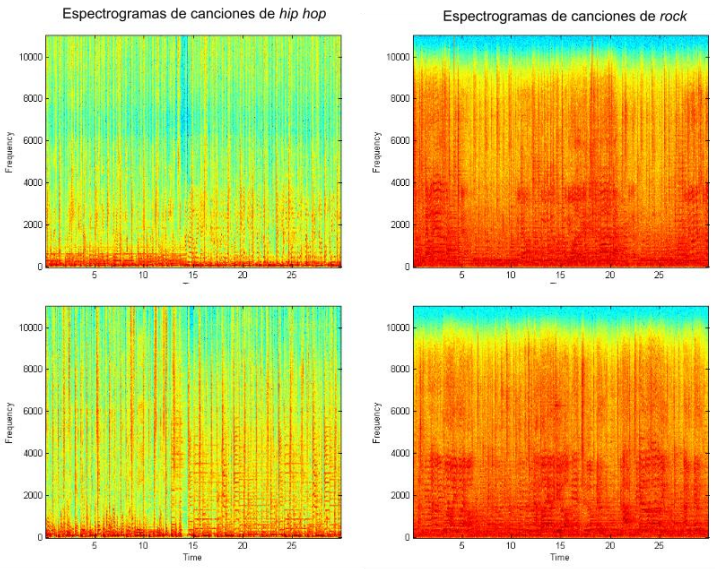


Figura 10, Deferencias y similitudes entre espectrogramas

4.3.1 MFCC

El método que se ha escogido para poder caracterizar esta información tímbrica, debido al extendido uso en el estado del arte [2], es el del uso de los *Mel-Frequency Cepstral Coefficients* (MFCCs, en adelante), explicados en la Sección 3.1.

Para el cálculo de dichos coeficientes se ha hecho uso del paquete de *software* libre de *LabROSA*[6], en concreto la función *melfcc.m*, usando los parámetros descritos en la **Tabla II**.

Parámetro	Frecuencia de muestreo [Hz]	Tiempo de ventana [s]	Tiempo de salto (*hoptime*) [s]	Número de coeficientes
Valor	22.050	0,02	0,01	21
Parámetro	Frecuencia máxima [Hz]	Número de bandas	Uso de energía	
Valor	11.025	22	True	

Tabla II, Elección de parámetros para la extracción de MFCCs

A continuación, se expone la explicación y justificación de los parámetros utilizados:

Frecuencia de muestreo: es la frecuencia máxima a la que están muestreadas los archivos de la base de datos *Marsyas*.

Tiempo de ventana: es la resolución de la ventana temporal donde se realiza la Transformada Discreta de Fourier (*Discrete Fourier Transform, DFT*) para el cálculo de los MFCCs. Se toma este valor, 20 ms, debido a que se consideró oportuno al evaluar el estudio realizado por Aucouturier y Pachet en [2]. La **Figura 11** muestra un criterio de precisión en función del tiempo de ventana para archivos de audio musicales. Se puede apreciar que existe un máximo en torno a 20 ms.

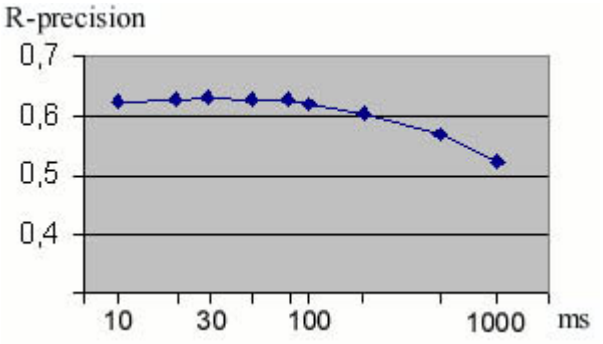


Figura 11, Influencia del tamaño de la ventana [2]

[6] LabROSA es un paquete *software* para extracción de características de señales de voz y audio. Se puede acceder en http://labrosa.ee.columbia.edu/matlab/rastamat/.

Tiempo de salto (*hoptime*): corresponde la diferencia entre el tiempo de una ventana con la siguiente ponderada por el porcentaje de solapamiento:

$$T_h = T_w(1 - R) \qquad (4.1)$$

Siendo T_h el tiempo de salto, T_w el tiempo de ventana y R el porcentaje de solapamiento. Debido a que tradicionalmente se utiliza en las Tecnologías del Habla un porcentaje de solapamiento R del 50%, se ha decidido hacer uso de ese valor en la **Ecuación 4.1** para la elección del tiempo de salto.

Número de coeficientes: número de coeficientes MFCC por ventana de trabajo. Se ha decidido tomar el valor N=21 debido a que se consideró oportuno al evaluar el estudio realizado por Aucouturier y Pachet en [2]. En la **Figura 12** se puede apreciar que, independientemente del número de componentes gaussianas del modelo utilizado (GMM), la gráfica muestra un máximo para el uso de 20 coeficientes. El valor de 21 corresponde con 20 coeficientes MFCC más 1 coeficiente de uso de energía.

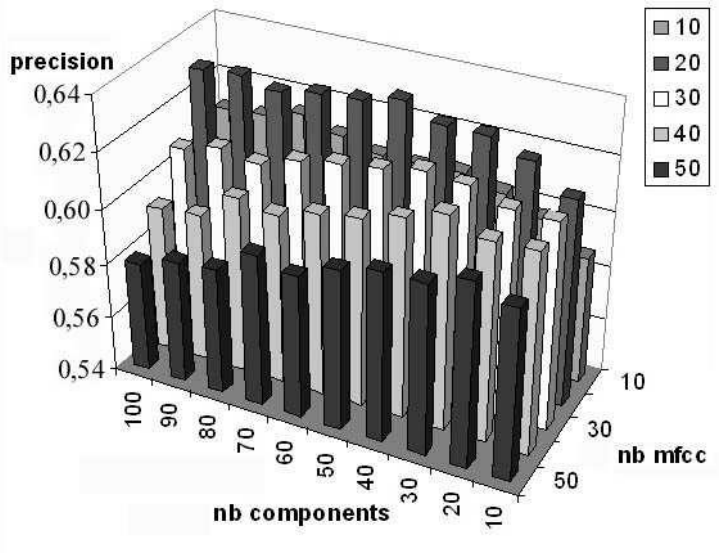


Figura 12, Relación entre componentes GMM y MFCCs [2]

Frecuencia máxima: es la frecuencia máxima del banco de filtros Mel para el cálculo de los MFCCs. Se ha tomado como valor la mitad de la frecuencia de muestreo, siendo perfectamente justificado haciendo uso de la teoría de muestreo de Nyquist:

$$F_{máx} = F_s/2 \qquad (4.2)$$

Siendo $F_{máx}$ la frecuencia máxima de la señal y F_s la frecuencia a la que ha sido muestreada la señal.

Número de bandas: número de bandas necesarias para la extracción de los MFCCs. Se ha decidido este valor haciendo uso de las recomendaciones de la función *melfcc.m* en la página web de *LabROSA*.

Uso de energía: en general, la energía de la ventana usada como una característica más del vector de características da buenos resultados en clasificación [2].

4.3.2 Otras técnicas

Además del uso de los MFCCs, en varias pruebas de este sistema se utilizan otras técnicas tales como el uso de la resta de la media cepstral, el cálculo de coeficientes delta y el uso de la tasa de cruces por cero, con vistas a obtener mejores resultados. Estas técnicas se explican a continuación:

Resta de la media cepstral (*Cepstral Mean Subtraction*, *CMS*): consiste en restarle al vector de características cepstrales, es decir, al de MFCCs, su vector media, con el fin de eliminar el efecto de un filtrado lineal que estuvieran afectando a toda la canción.

Coeficientes delta (*Delta-Spectral Cepstral Coefficients)*: los coeficientes delta miden la variación de los vectores de características MFCC dentro de una ventana de vectores de características. El tamaño de la ventana utilizado ha sido $W=9$, debido a que debe ser impar y medianamente grande como para tener una estimación de cómo varía la señal a lo largo del tiempo (parámetro necesario de la función *deltas.m* del *software* de *LabROSA*). Utilizando ese tamaño de ventana, se obtiene que, para aplicar la **Ecuación 4.3**, $B = \frac{W-1}{2}$.

$$d(\tau) = \frac{\sum_{k=1}^{B} k(c(\tau+k) - c(\tau-k))}{2\sum_{k=1}^{W/2} k^2} \qquad (4.3)$$

Donde $c(\tau)$ es el coeficiente MFCC en el instante τ.

Tasa de cruces por cero (*Zero-Crossing Rate*, *ZCR*): con vistas a romper el límite en el rendimiento que se produce al trabajar únicamente con información de tipo cepstral, se decide utilizar la tasa de cruces por cero como una característica más, para evaluar si el sistema mejora cuando se utiliza información no cepstral. La tasa de cruces por cero se utiliza en cada ventana, con un solapamiento igual al que se utiliza para calcular los MFCCs. Se aplica la siguiente ecuación para su cálculo:

$$ZCR(x) = \frac{\sum count(x)}{L_w} \qquad (4.4)$$

Siendo x las muestras de la ventana, $count(x)$ una función que cuenta las veces que dos muestras consecutivas de x cambian de signo y L_w el número de muestras de la ventana.

4.4 Los modelos

Los modelos son, en este caso, modelos probabilísticos utilizados para representar a los géneros musicales. Existen dos fases independientes que tienen relación directa con los modelos: el primero, la fase de entrenamiento, que permite diseñar y ajustar un modelo haciendo uso de los datos de entrenamiento, y, una vez esté constituido el modelo, la segunda fase es la fase de test, que permite evaluar el rendimiento del modelo, haciendo uso de los datos de test, ya que el clasificador diseñado es un sistema de tipo supervisado.

Los dos modelos utilizados en este sistema son el Modelo de Mezclas de Gaussianas (*Gaussian Mixture Model*, GMM a partir de ahora) y el Modelo Universal (*Universal Background Model*, UBM a partir de ahora). Ambos modelos, como bien se ha explicado en la Sección 3.2, utilizan un conjunto de *M* funciones gaussianas que, mediante un método iterativo y una función de coste (*K-means* para la inicialización, *Expectation Maximization* para el entrenamiento) se ajustan a los datos para producir el menor coste, es decir, la mínima desviación entre la función densidad de probabilidad de la mezcla de las gaussianas y la de los datos [11][12].

En ambos modelos se ha decidido tomar un conjunto de *M*=64 gaussianas, debido a que se consideró oportuno tras evaluar el estudio de Aucouturier y Pachet en [2], y como se puede observar en la **Figura 12** en la Sección 4.3.1, y a que era un número potencia de 2, que en general suele ser beneficioso en términos de escalabilidad.

El resultado de los dos modelos son tres parámetros por gaussiana:

Vector de medias (μ): es un vector de tantas componentes como parámetros tiene el vector de características *N*. Corresponde con la media de cada gaussiana.

Matriz/Vector de covarianza (Σ/σ, respectivamente): el primero corresponde con la matriz de covarianza de la gaussiana, de tamaño *NxN*. El segundo corresponde el vector diagonal de la matriz de covarianza, de tamaño *N*. En términos generales, si no se especifica que se utiliza una matriz de covarianza diagonal se hará uso de Σ, la matriz de covarianza. En el caso en el que se haga uso de una matriz de covarianza diagonal se entenderá esta matriz como:

$$\Sigma = \sigma \mathrm{I} \qquad (4.5)$$

Siendo I la matriz unidad de tamaño *NxN*.

Peso (*w*): corresponde al valor de ponderación por el que va multiplicado una gaussiana. Es un escalar. La suma de los pesos de las *M* gaussianas debe sumar 1 para que un GMM sea una función de densidad de probabilidad.

Debido a que se están utilizando *M*=64 gaussianas, se denota como **μ** a la matriz *MxN* de centros, **Σ** a la hipermatriz *NxNxM* de covarianzas, y **w** al vector de *M* pesos. Por lo tanto, con estos tres parámetros, **μ**, **Σ** y **w**, queda totalmente definido un modelo.

Es importante tener en cuenta que los modelos hacen uso de las características extraídas en la etapa de extracción de características, y no directamente de la información de forma de onda, por los motivos anteriormente expuestos en la Sección 4.3.

El *software* para implementar la fase de entrenamiento y de test de ambos modelos ha hecho uso de las funciones del paquete libre *Netlab*[7]. Se han utilizado las funciones *gmm.m* para inicializar la estructura de mezclas, la función *gmminit.m* para inicializar los parámetros mediante el algoritmo de *K-means* —haciendo uso de una sola iteración—, la función *gmmem.m* para ajustar los parámetros a los datos —haciendo uso de cinco iteraciones—. La función utilizada para la fase de test del paquete libre *Netlab* ha sido *gmmprob.m*, que calcula la verosimilitud de un conjunto de vectores de características dado un modelo.

[7] Netlab es un paquete de *software* libre utilizado en el cálculo de GMM. Se puede acceder en http://www.mathworks.com/matlabcentral/fileexchange/2654-netlab .

4.4.1 GMM con ML

Como queda expuesto en la Sección 3.2, GMM [11] es un modelo matemático que define una función densidad de probabilidad a partir de mezclas de gaussianas. En este apartado se explica la aplicación de este método al problema concreto de clasificación de género y las variables que se han modificado en las diferentes pruebas realizadas.

Este modelo, GMM, se suele utilizar con frecuencia junto con el criterio de máxima verosimilitud (*Maximum Likelihood*, ML en adelante), por eso se hace referencia a él como GMM-ML.

Fase de entrenamiento

Durante la fase de entrenamiento, se dividen los datos de entrenamiento por género y se entrena un modelo por cada género, como representa la **Figura 13**.

Se utilizan las características extraídas de los datos de entrenamiento y se entrenan los vectores de media, matrices de covarianza y pesos por cada gaussiana de los 10 modelos de género mediante un proceso iterativo, con el fin de minimizar el error, como se explica en la Sección 3.2.

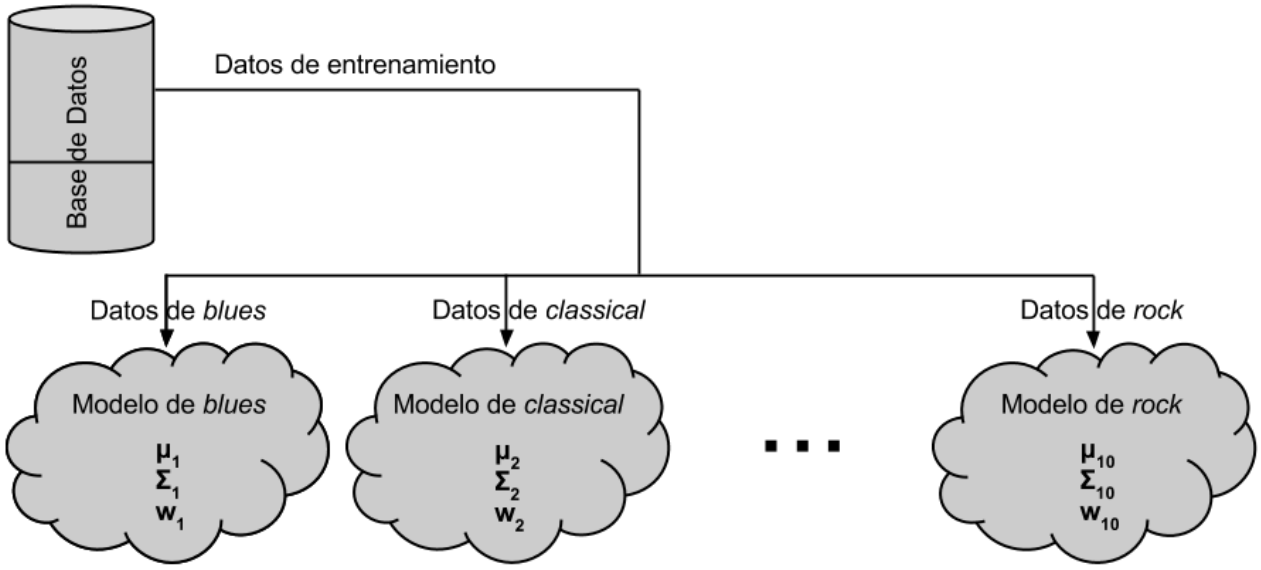


Figura 13, Entrenamiento GMM

El parámetro que se ha modificado en las diferentes pruebas es el tipo de matriz de covarianza que se desea entrenar que, o bien es completa o bien es diagonal. Como se explica en la Sección 3.2, el uso de matrices diagonales minimizan el tiempo de cómputo (se calculan *NxM* valores por modelo de género), mientras que si se entrena un modelo con matrices de covarianza completas se consigue representar mejor la función densidad de probabilidad de cada género, pero el coste computacional crece significativamente (se calculan *NxNxM* valores por modelo de género).

Fase de Test

Durante la fase de test, se calcula la probabilidad de que una canción pertenezca a un género en concreto. Para ello se realiza el siguiente proceso:

1. Se toma una canción del conjunto de datos de test, donde se han extraído previamente los vectores de características.
2. Por cada vector de características se calcula la verosimilitud (**Ecuación 3.5**) de que ese vector pertenezca a un modelo de género dado. Este paso se repite para todos los vectores de características de una canción dada.
3. Se halla el logaritmo de todas las verosimilitudes calculadas en el paso anterior y se realiza la suma de ellas, como indica la **Ecuación 4.6**, obteniendo así la puntuación o *score* (la **Ecuación 4.6** se deriva de transformar al dominio logaritmo la **Ecuación 3.7**). Posteriormente, se normaliza el resultado respecto al número de vectores de características extraídos en esa canción. Cuanto mayor sea la puntuación, más se parece una canción a un género musical.
4. Se realizan los pasos 2 y 3 para todos los géneros.
5. Se realizan los pasos 1 al 4 para todas las canciones del conjunto de datos de test.

$$\text{score} = \sum_{t=1}^{T} \log(p(v_t|\lambda_j)) \,/T \qquad (4.6)$$

En la **Ecuación 4.6**, se define $p(v_t|\lambda)$ como la verosimilitud de todos los vectores de características v_t, dado el modelo λ_j. T es el número de vectores de entrenamiento de esa canción.

Por lo tanto, por cada canción tendremos 10 puntuaciones, una para cada género. Como se explica en la Sección 3.2, se utiliza la suma del logaritmo de la verosimilitud de las características del audio dado el modelo de género, ya que se supone independencia entre todos los vectores de características.

Para asignar una canción a un determinado género se elegirá aquél con la mayor puntuación.

4.4.2 UBM con MAP

Como se ha explicado en la Sección 3.2, UBM [11] se basa en entrenar mediante un GMM un modelo general o universal al que podríamos llamar *modelo de música*, y adaptar, mediante las canciones de entrenamiento de cada género, distintos *modelo de género*. La adaptación se ha decidido hacer mediante el método MAP (*Maximum A Posteriori*) [11], frente a MLLR (*Maximum Likelihood Linear Regression*), debido a su uso extensivo en el área de reconocimiento de locutor.

Fase de entrenamiento

Como muestra la **Figura 14**, en primer lugar es necesario entrenar un modelo universal con todos los datos de entrenamiento. Posteriormente, utilizando los datos de entrenamiento de cada género se hace una adaptación MAP, como se explica en la Sección 3.2.

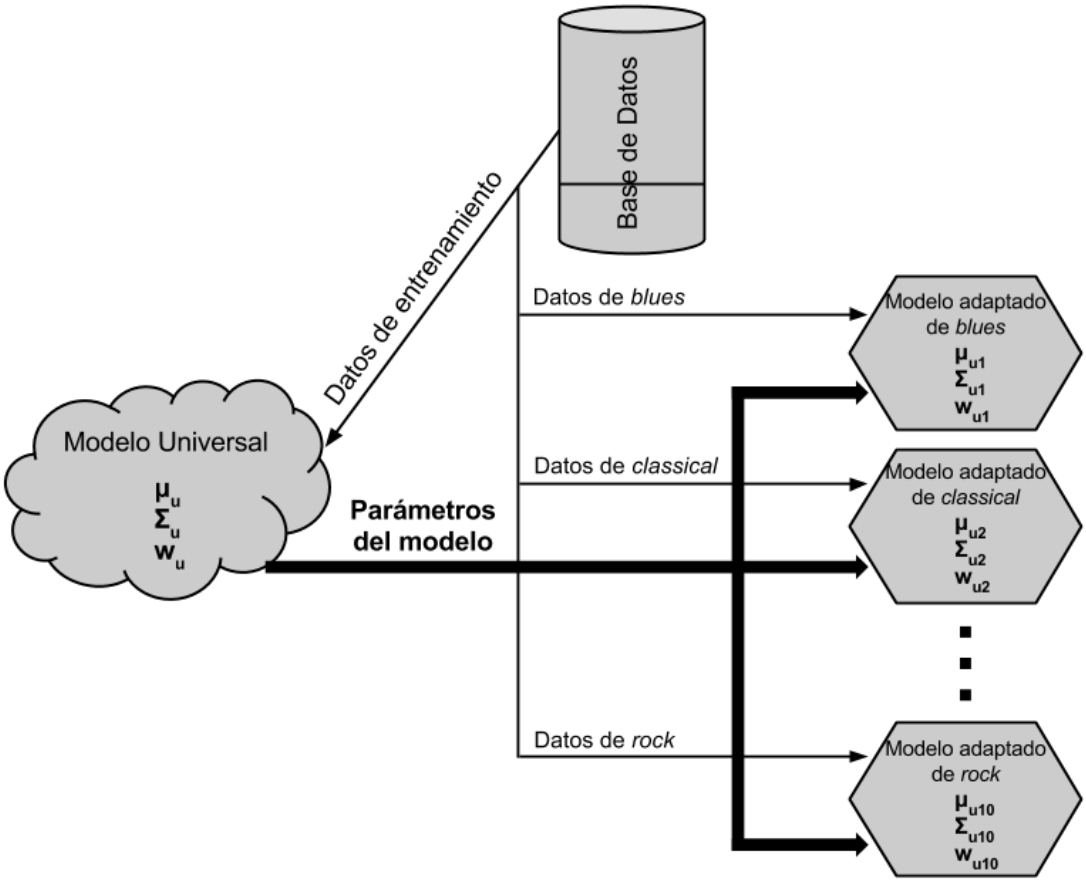


Figura 14, Entrenamiento UBM

Debido a problemas de potencia computacional, para generar el modelo universal se han tenido que diezmar los datos de entrenamiento por un factor 2, es decir, se toma 1 de cada 2 vectores de características de los datos de entrenamiento para entrenar un modelo.

Al igual que en el modelo GMM-ML, el parámetro que se puede modificar para la generación del UBM es el tipo de covarianza utilizada, que bien puede ser completa, o bien diagonal.

Para la adaptación MAP se pueden modificar los parámetros de adaptación de medias, de pesos y de covarianzas, siendo este último caso posible sólo si la matriz de covarianza es diagonal, debido al alto coste computacional que supone este cálculo.

Fase de Test

El cálculo de las puntuaciones (*scores*) para el modelo GMM-UBM es similar al caso GMM-ML. La diferencia principal radica en que, además de calcular la verosimilitud de que una canción pertenezca a los 10 diferentes géneros (paso 4 de la *Fase de Test* del modelo GMM-ML, Sección 4.4.1) se calcula también la verosimilitud de que una canción pertenezca al modelo universal. Posteriormente, una vez calculadas las puntuaciones resultantes de enfrentar cada canción a los 10 géneros y al modelo universal, se normaliza la puntuación de cada uno de los 10 géneros respecto a la puntuación del modelo universal, restando al valor del logaritmo de la puntuación obtenida por los modelos de género el logaritmo de la puntuación del modelo universal, como muestra la **Ecuación 4.7**, derivada de combinar las **Ecuaciones 3.7** y **3.8** y transformar el resultado al dominio logaritmo:

$$\text{score}_{\text{UBM}} = \sum_{i} \log\left(p(v_i|\lambda_j)\right) - \sum_{i} \log\left(p(v_i|\lambda_U)\right) \qquad (4.7)$$

Donde, al igual que en la **Ecuación 4.6**, $p(v_i|\lambda)$ es la verosimilitud de todos los vectores de características v_i, dado el modelo λ_j y donde λ_U es el modelo universal.

Al igual que en el modelo GMM-ML, se toma la mayor puntuación de entre el conjunto de las 10 puntuaciones normalizadas resultantes.

4.5 La evaluación de resultados

La evaluación de resultados es la última etapa del clasificador y en ella se busca un criterio de bondad del clasificador para considerar si una mejora es positiva o negativa. El criterio utilizado es el porcentaje de aciertos de clasificación. Si una canción de test se clasifica en el género al que pertenece, se considerará acierto. En cualquier otro caso se considerará fallo. Se ha considerado que el porcentaje de aciertos es una medida adecuada porque hay equilibrio entre las diferentes clases, ya que hay el mismo número de puntuaciones de cada género.

Además del porcentaje de aciertos de clasificación, se han querido también evaluar los datos en una matriz de confusión, **Figura 15**. La matriz de confusión es una matriz cuadrada de *10x10* donde en el eje de abscisas se observan los 10 géneros reales y en el eje de ordenadas aparecen los 10 géneros asignados. De esta manera, se puede evaluar si la clasificación es correcta si la diagonal tiene alta puntuación, cuáles son los géneros que más se confunden, por cuáles se confunden, etcétera.

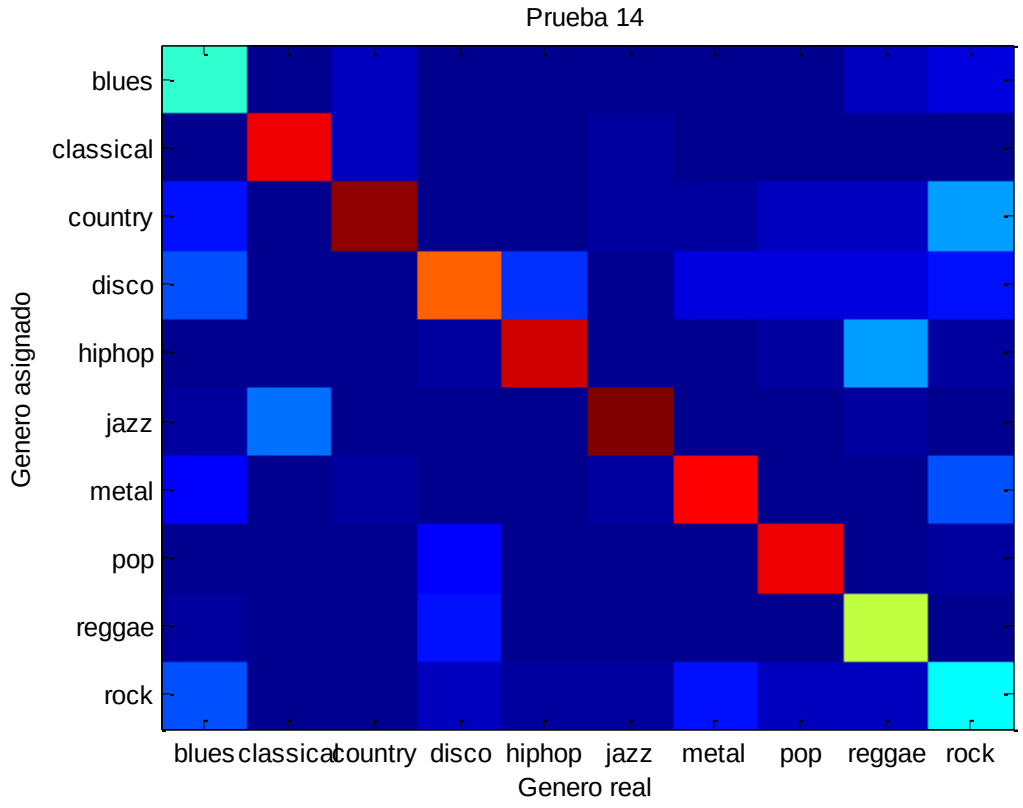


Figura 15, Ejemplo de una matriz de confusión

SECCIÓN 5: PRUEBAS Y RESULTADOS

En esta sección se explican los parámetros que se han modificado a lo largo de las diferentes pruebas realizadas. Se toma como resultado a evaluar y comparar el porcentaje de aciertos de clasificación. En la **Figura 16** se muestran todas las matrices de confusión de las pruebas realizadas.

Al final de esta sección se hará un análisis de resultados de todas las pruebas de manera conjunta, que servirá como base a la hora de extraer las conclusiones. En la Sección 6 se extraerán las conclusiones interpretando el análisis realizado en esta sección.

Los parámetros que varían de una prueba a otra son los siguientes:

- **Modelo**: tipo de modelo utilizado. Puede ser GMM (GMM con ML) o UBM (GMM con UBM y adaptación MAP). Explicados en la Sección 4.3.
- **CMS**: Uso o no de la técnica de Resta de Media Cepstral. Explicado en la Sección 4.2.2.
- **Deltas:** Uso o no de la característica Coeficientes Delta. Explicado en la Sección 4.2.2.
- **ZCR:** Uso o no de la característica Tasa de Cruces por Cero. Explicado en la Sección 4.2.2.
- **N:** Número de coeficientes de los vectores de características.
- **CovType**: tipo de covarianza. Puede ser diagonal (*diag*) o completa (*full*). Explicado en la Sección 3.2.
- **Adapt**: parámetros de adaptación MAP. Pueden ser varios, a saber vectores de media (*m*), pesos (*w*) y matrices de covarianzas (*c*). Explicado en la Sección 3.2.
- **Norm:** Uso o no de normalización de puntuaciones. Puede usarse T-Norm (*T*) o Z-Norm (*Z*). La normalización se utiliza una vez extraídas las puntuaciones, antes de evaluar los resultados, como se explica en la Sección 3.3.

Prueba	Modelo	CMS	Deltas	ZCR	N	CovType	Adapt	Norm	%Hit
1	GMM	-	-	-	21	diag	-	-	51,50
2	UBM	-	-	-	21	diag	m	-	40,25
3	GMM	Sí	-	-	21	diag	-	-	61,00
4	UBM	Sí	-	-	21	diag	m	-	58,50
5	GMM	Sí	-	-	21	full	-	-	63,50
6	UBM	Sí	-		21	full	m	-	59,00
7	UBM	Sí	-	-	21	full	m, w	-	57,75
8	UBM	Sí	-	-	21	diag	m, w	-	60,00
9	UBM	Sí	-	-	21	diag	m, w, c	-	56,75
10	GMM	Sí	Sí	-	42	diag	-	-	67,50
11	UBM	Sí	Sí	-	42	diag	m, w	-	**65,25**
12	GMM	Sí	Sí	Sí	43	diag	-	-	61,50
13	UBM	Sí	Sí	Sí	43	diag	m, w	-	62,50
14	GMM	Sí	-	-	21	diag	-	T	**68,75**
15	UBM	Sí	-	-	21	diag	m, w	T	60,00
16	GMM	Sí	-	-	21	diag	-	Z	14,50
17	UBM	Sí	-	-	21	diag	m, w	Z	50,75
18	GMM	Sí	Sí	-	42	diag	-	T	68,75
19	UBM	Sí	Sí	-	42	diag	m, w	T	65,25
20	GMM	Sí	Sí	-	42	diag	-	Z	10,00
21	UBM	Sí	Sí	-	42	diag	m, w	Z	46,25
22	GMM	Sí	Sí	Sí	43	diag	-	T	68,75
23	UBM	Sí	Sí	Sí	43	diag	m, w	T	62,50
24	GMM	Sí	Sí	Sí	43	diag	-	Z	10,00
25	UBM	Sí	Sí	Sí	43	diag	m, w	Z	43,25

Tabla III, Tabla de pruebas y resultados

Las **Figuras 17**, **18** y **19** muestran los porcentajes de aciertos en función del modelo de una forma gráfica. La **Figura 17** representa gráficamente el porcentaje de aciertos de todos los modelos. Las **Figuras 18** y **19** muestran por separado los porcentajes de aciertos de las pruebas de los modelos GMM-ML y GMM-UBM respectivamente.

Análisis de resultados

Como se puede observar en la **Tabla III**, el mejor modelo, la Prueba 14, tiene un porcentaje de aciertos del 68,75% y consta de un modelo GMM-ML de matriz diagonal con 21 características, con Resta de la Media Cepstral (CMS) y Normalización de tipo T.

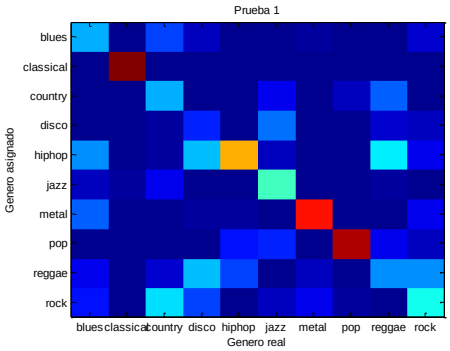
Ranking	Prueba	% Aciertos	Modelo
1	14	68,75	GMM
2	18	68,75	GMM
3	22	68,75	GMM
4	10	67,50	GMM
5	11	65,25	UBM
6	19	65,25	UBM
7	5	63,50	GMM
8	13	62,50	UBM
9	23	62,50	UBM
10	12	61,50	GMM

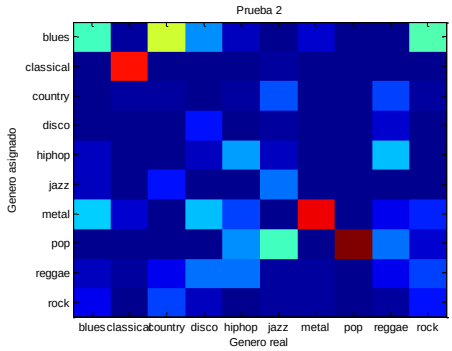
Tabla IV, Tabla de las 10 mejores pruebas

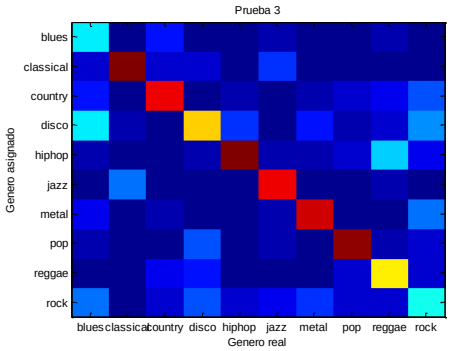
El mejor modelo GMM-UBM es la Prueba 11, con porcentaje de aciertos del 65,25%, con matriz diagonal, adaptación de medias y pesos, con 42 coeficientes, Resta de la Media Cepstral (CMS) y Coeficientes Delta.

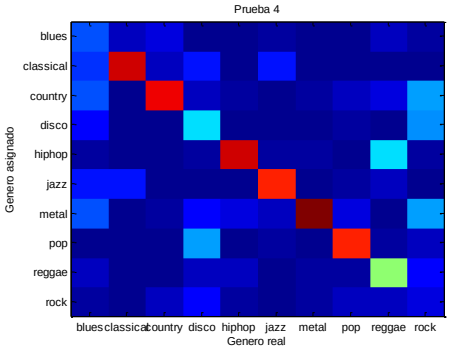
En la **Tabla IV** se muestra el *ranking* de las 10 mejores puntuaciones, la prueba a la que corresponde y el tipo de modelo. A partir de esta tabla y su relación con la Tabla III se pueden observar lo siguiente:

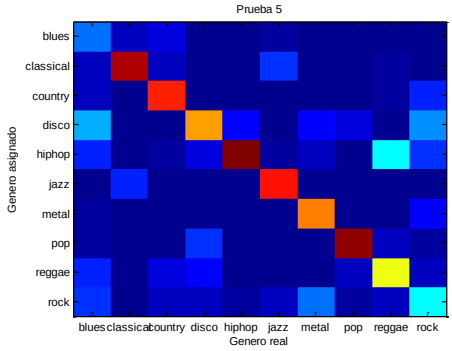
- Los tres mejores modelos son de tipo GMM-ML y el porcentaje de aciertos es el mismo: 68,75%, además de que en las tres pruebas se ha clasificado igual, es decir, se han producido los mismos errores de clasificación. También existe una característica común entre estos tres modelos y es que han sido normalizados mediante la Normalización T. En la **Figura 21** se muestran las distribuciones (mediante diagramas de cajas) de las puntaciones de las canciones de cada género respecto a todos los géneros habiendo utilizado la Normalización T (Prueba 14). Se puede observar, haciendo una comparación con la **Figura 20** (que muestra los diagramas de cajas en la Prueba 3), que existe un realce en las puntuaciones de un género para las canciones del mismo género.
- El modelo GMM-UBM no aparece hasta las posiciones 5 y 6 del *ranking*, con un porcentaje de aciertos de 65,25%. La diferencia entre las pruebas 11 y 19 es el uso de la Normalización T, utilizada en la Prueba 19, por los demás, los parámetros utilizados son los mismos. Como se ve también en la **Tabla III** los resultados entre las parejas de Pruebas 11 y 19, 8 y 15 y 13 y 23 nos da a entender que la Normalización T no varía el porcentaje de aciertos en UBM al utilizarlo.
- A la hora de realizar la Normalización Z, los datos escogidos para realizar la normalización influyen considerablemente en el resultado final. En la **Ilustración 22** se muestran los diagramas de cajas de las puntuaciones de las canciones de cada género respecto a todos los géneros, habiendo utilizado la Normalización Z (Prueba 16).
- El uso de una matriz de covarianza completa no es muy significativo a la hora de mejorar la eficacia de clasificación (Prueba 5 frente a 3, Prueba 6 frente a 4 y Prueba 7 frente a 8). Sí podríamos decir que la adaptación de matrices de covarianzas no hace mejorar el sistema (Prueba 9 frente a 8).

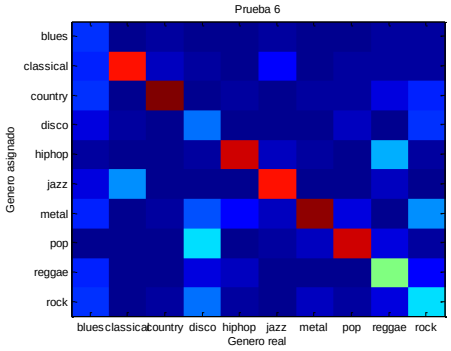

Prueba 1
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

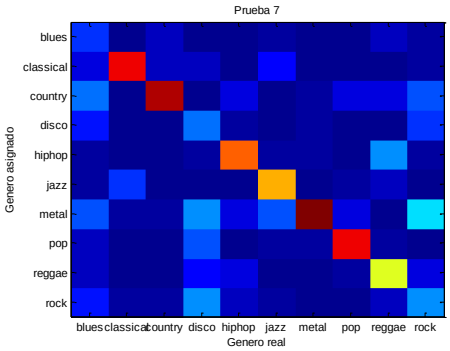

Prueba 2
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

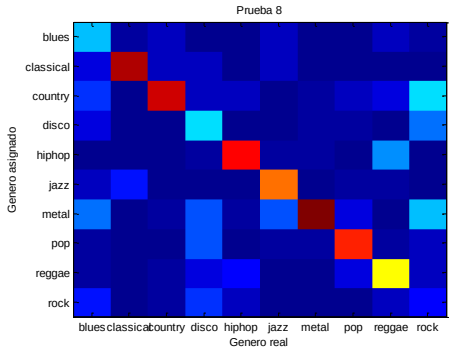

Prueba 3
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

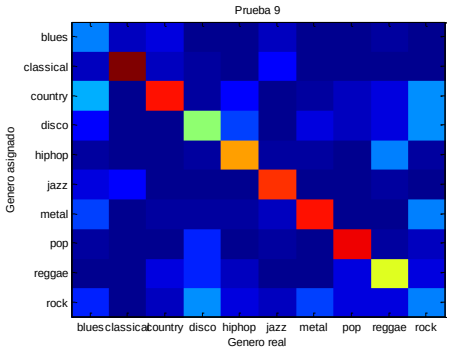

Prueba 4
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

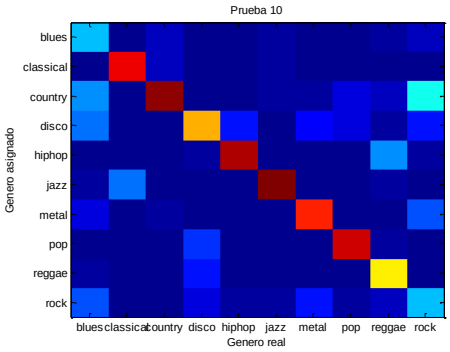

Prueba 5
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

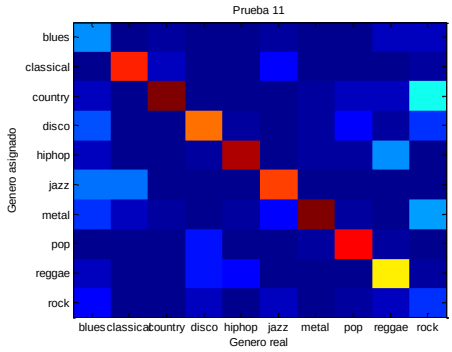

Prueba 6
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

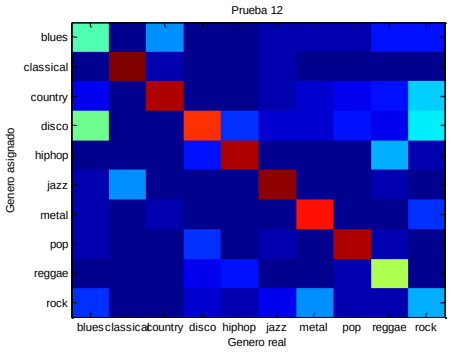

Prueba 7
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

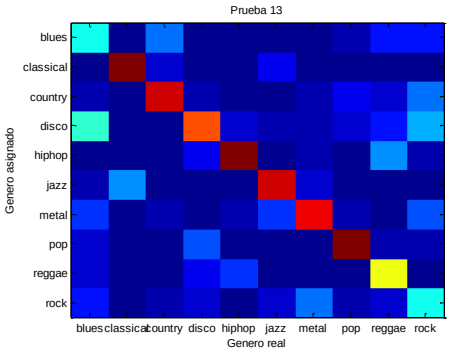

Prueba 8
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

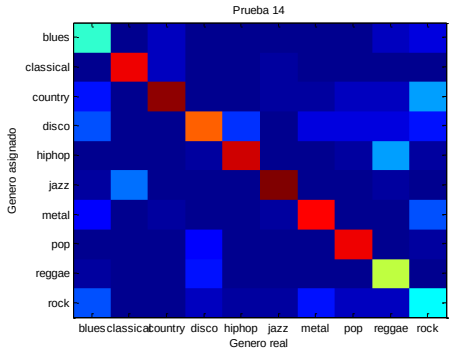

Prueba 9
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

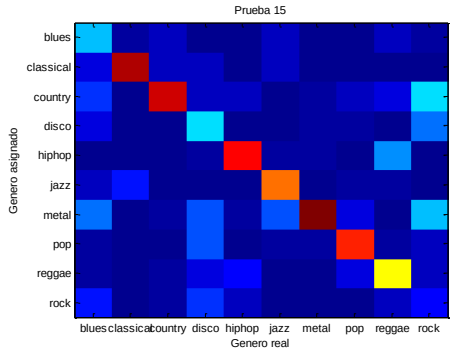

Prueba 10
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

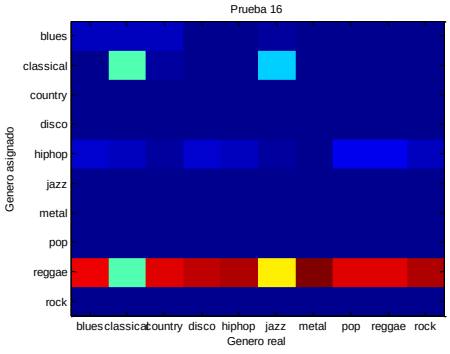

Prueba 11
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

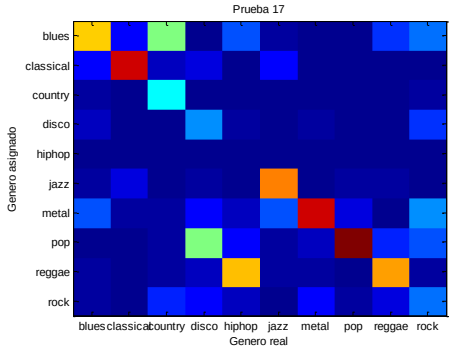

Prueba 12
Genero asignado
Genero real
blues classical country disco hiphop jazz metal pop reggae rock

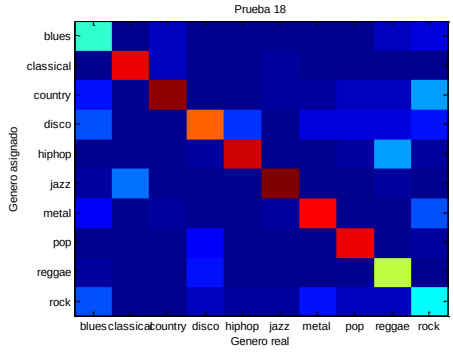

Prueba 13
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real

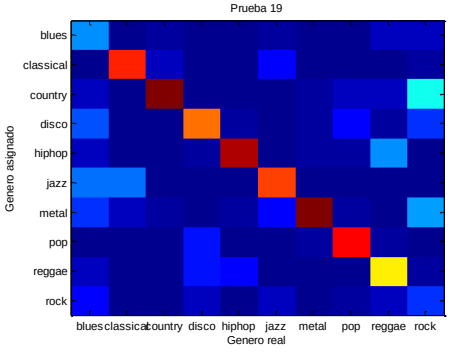

Prueba 14
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real

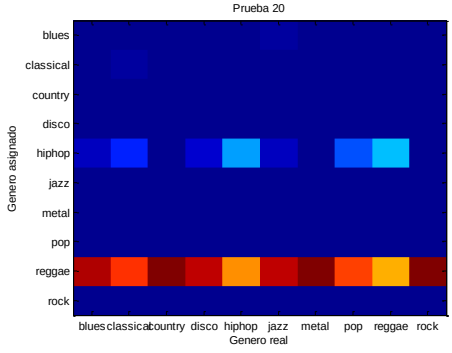

Prueba 15
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real

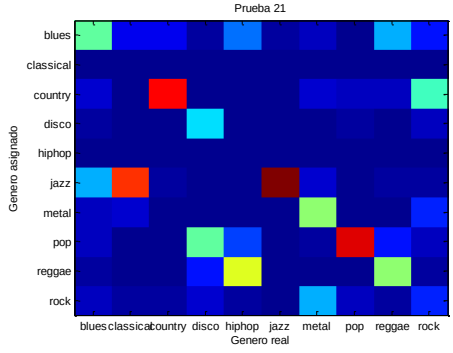

Prueba 16
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real

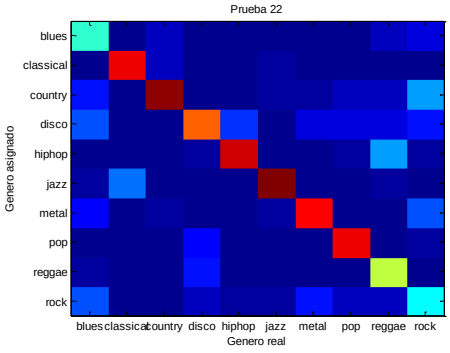

Prueba 17
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real

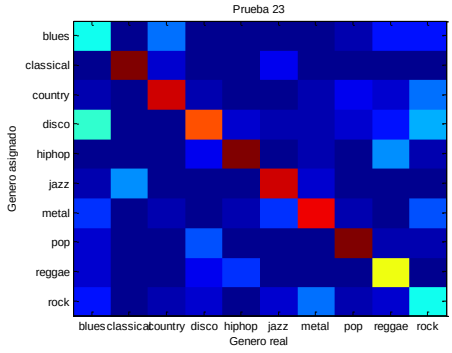

Prueba 18
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real

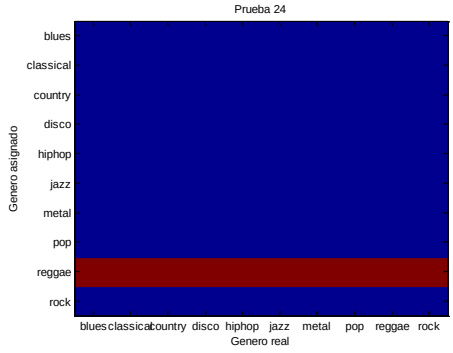

Prueba 19
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real


Prueba 20
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real


Prueba 21
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real


Prueba 22
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real


Prueba 23
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real


Prueba 24
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Genero asignado
blues classical country disco hiphop jazz metal pop reggae rock
Genero real

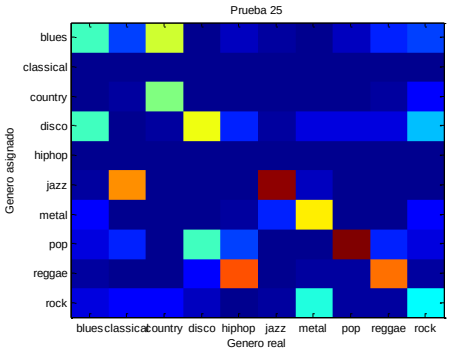


Figura 16, Matrices de confusión

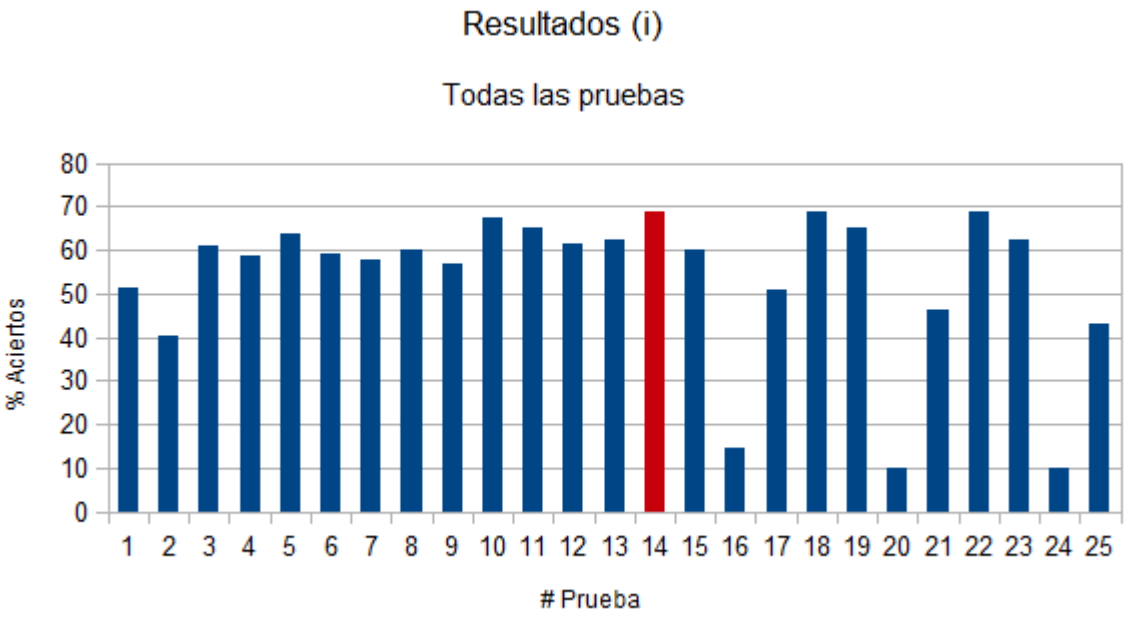


Figura 17, Diagrama de barras de los resultados de todas las pruebas realizadas

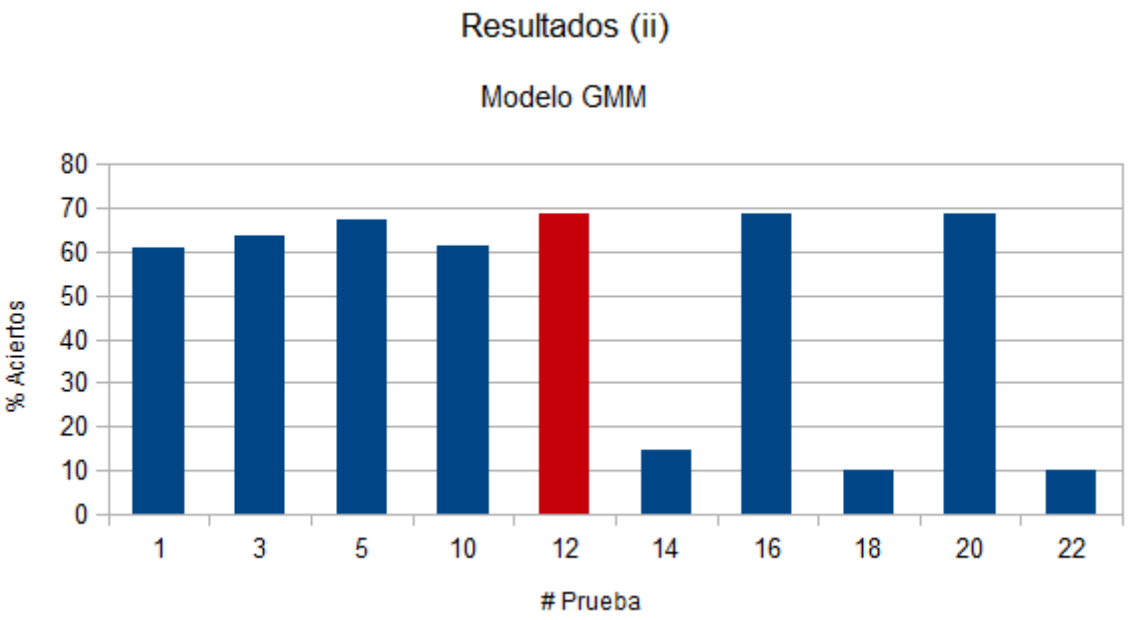


Figura 18, Diagrama de barras de los resultados de las pruebas con modelo GMM

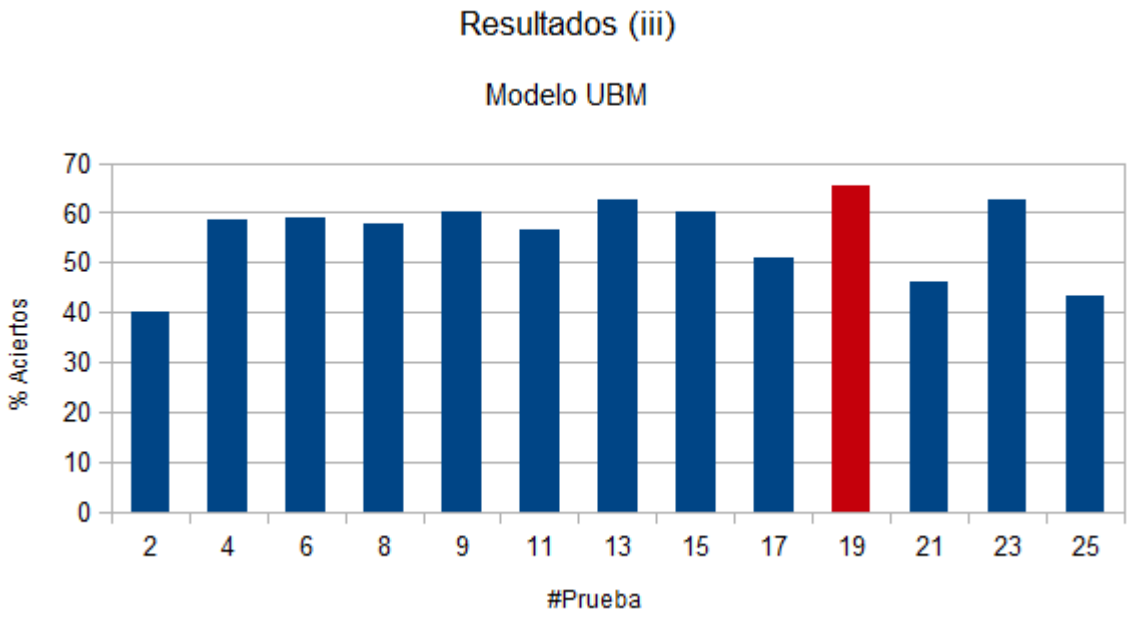


Figura 19, Diagrama de barras de los resultados de las pruebas con modelo UBM

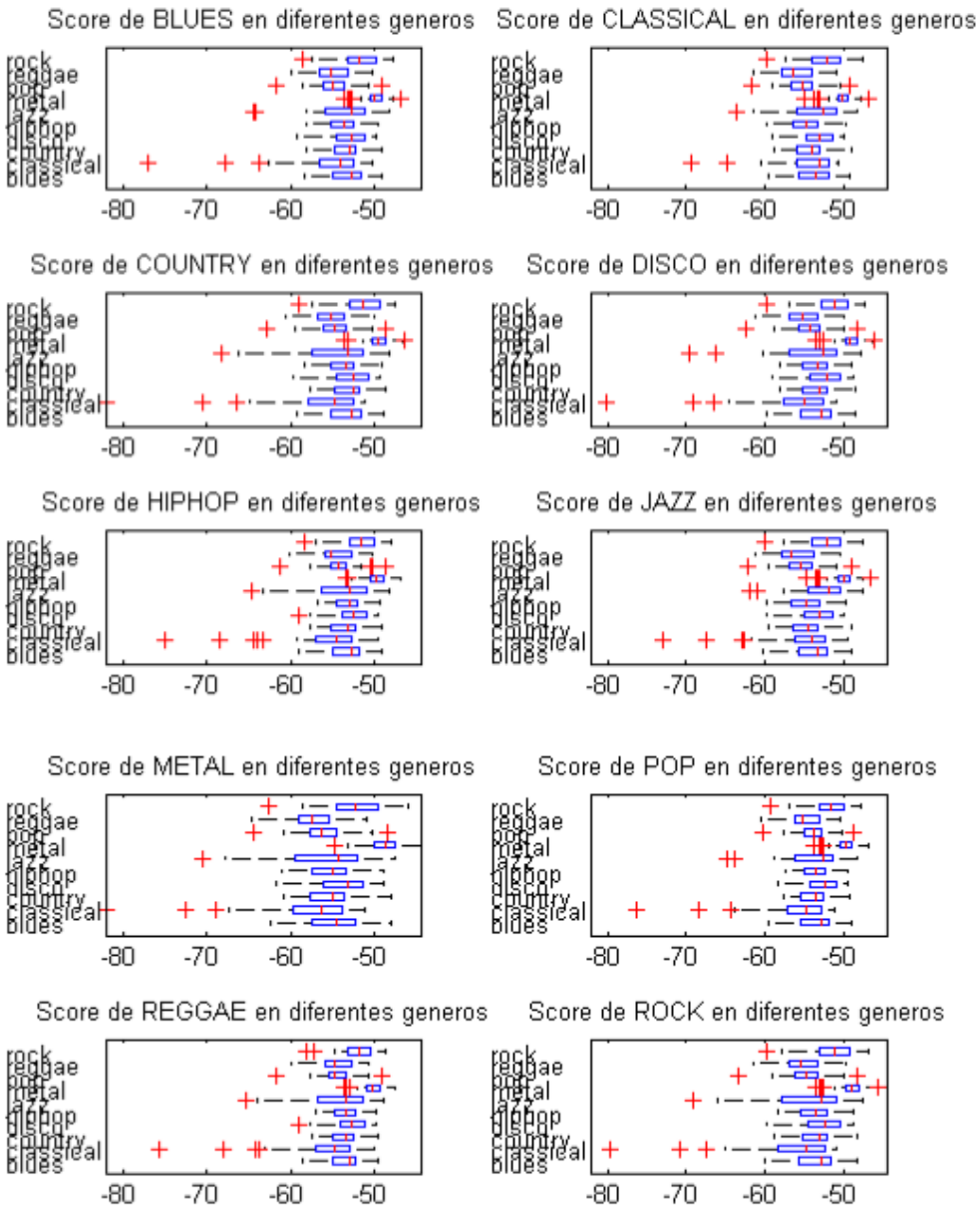


Figura 20, Histograma de puntuaciones - Prueba 3

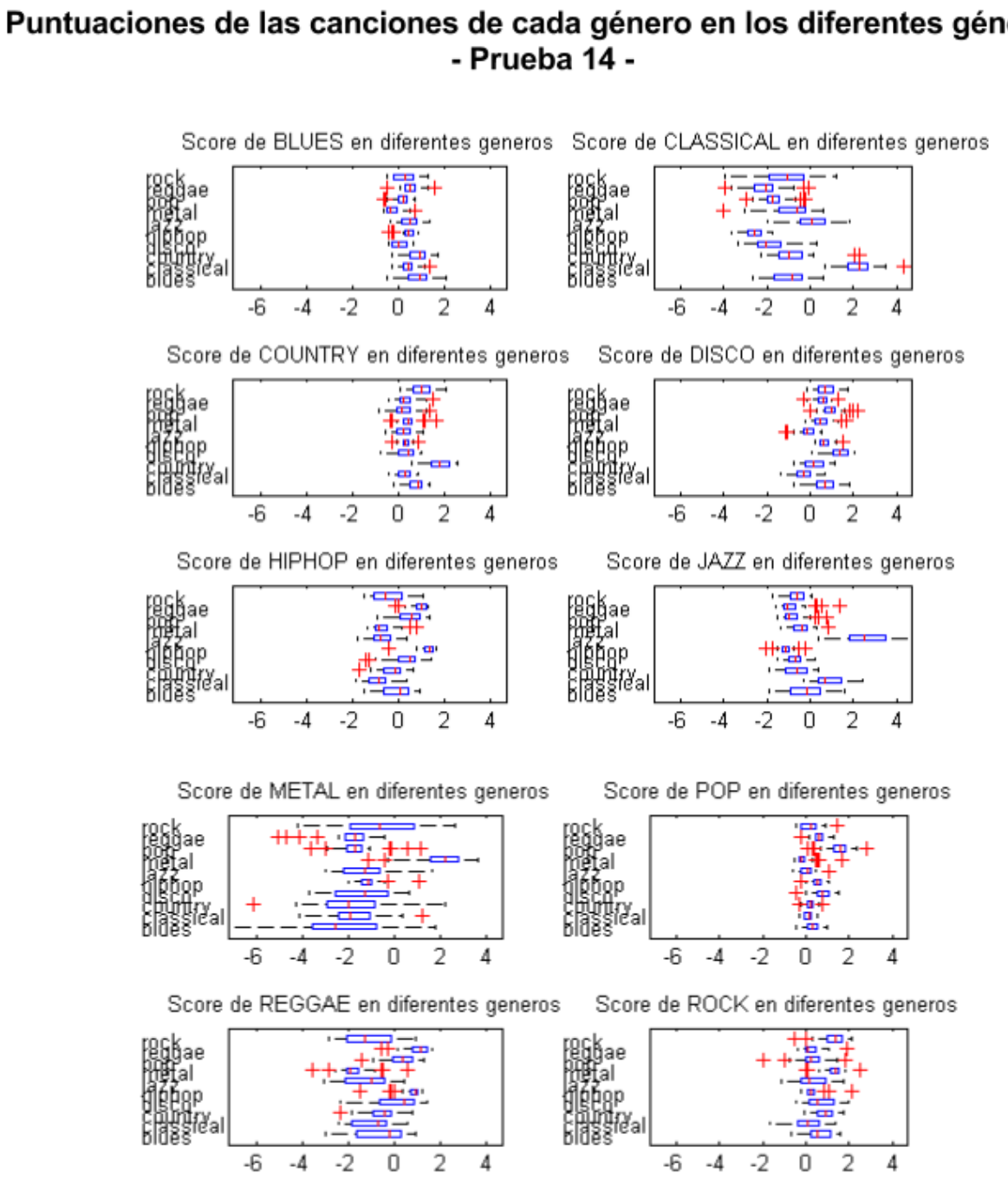


Figura 21, Histograma de puntuaciones - Prueba 14

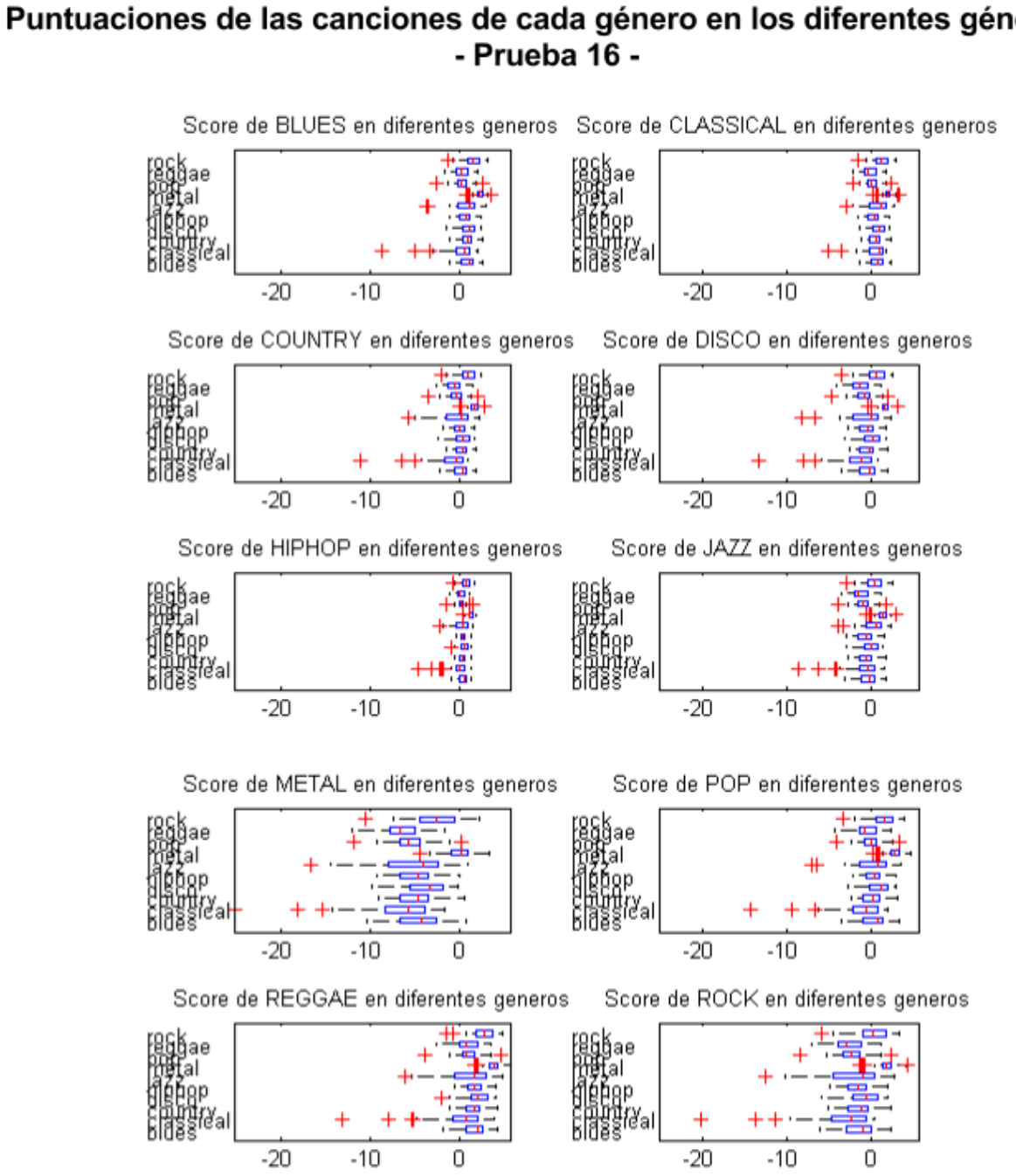


Figura 22, Histograma de puntuaciones - Prueba 16

SECCIÓN 6: CONCLUSIONES Y TRABAJO FUTURO

6.1 Conclusiones

A la vista del análisis de resultados hecho en la sección anterior se pueden extraer las siguientes conclusiones:

- Que los cuatro primeros mejores modelos correspondan con que sean de tipo GMM-ML puede deberse a la falta de datos a la hora de entrenar los modelos, o a que sean modelos muy ajustados. UBM necesita una gran cantidad de datos de mayor variabilidad para que funcione de manera más eficiente que el modelo GMM-ML [11] [12].
- La Resta de la Media Cepstral es una técnica que hace mejorar los resultados de clasificación, al igual que el uso de Coeficientes Delta. Como se observa en la **Tabla III**, al aplicar la técnica del CMS consiguen 10 puntos de mejora en el modelo GMM-ML y hasta 18 en GMM-UBM, y al usar las Deltas, 5 puntos en el modelo GMM-ML y 6 en el modelo GMM-UBM. Esto es claramente porque muchos efectos de producción musical que pueden distorsionar el espectro son lineales, y por tanto se compensan muy bien con CMS.
- La Normalización T hace mejorar el modelo GMM-ML, pero se mantiene invariante para el modelo GMM-UBM. Esto se debe a que GMM-UBM consiste en normalizar las puntuaciones de los géneros respecto al modelo universal y, por ello, a las puntuaciones del modelo GMM-UBM ya se les ha aplicado una normalización. También esto nos da a entender que el modelo GMM-ML necesita normalización de puntuaciones.
- GMM-UBM es un modelo más robusto que el modelo GMM-ML, ya que al aplicar la Normalización Z, la puntuación del modelo GMM-UBM se ha mantenido en valores más altos que en el modelo GMM-ML.
- El uso de matrices de covarianza completa no hace mejorar el sistema probablemente debido al hecho de que los modelos utilizados son muy complejos, es decir, con muchas mezclas, y modelan bien la correlación local entre muestras al utilizar matrices de covarianza diagonales. GMM-UBM, por su lado, modela suficientemente bien cuando se ajustan las medias y pesos como para ajustar además sus matrices de covarianza con la estimación MAP, lo que hace que los resultados empeoren.

- El hecho de que el porcentaje de clasificación sea 68,75% en los tres mejores modelos y que los errores de clasificación sean los mismos puede estar relacionado de manera muy directa con la base de datos utilizada, es decir, se llega a un límite en la clasificación debido a la naturaleza de los datos utilizados tanto en las fases de entrenamiento y test de los modelos.
- El uso de la Tasa de Cruces por Cero no hace mejorar los resultados. Puede deberse este hecho a las condiciones de grabación de las canciones de la base de datos, que pueden ser ruidosas y por ello provocan que esta característica no sea muy discriminativa. También puede deberse al hecho de que la ZCR a corto plazo se puede considerar como una medida poco precisa de la frecuencia instantánea, es decir, tiene una relación con el *pitch*. Puede ser que el *pitch* por sí solo no sea muy discriminativo en los géneros musicales y por ello haga que el clasificador se confunda.
- Los errores de clasificación en las Pruebas 14, 18 y 22, aquellas con mejores resultados, son musicalmente bastante coherentes. Como se puede observar en la **Figura 23**, los errores de clasificación del género *Blues* aparecen en los géneros *Country*, *Disco*, *Metal* y *Rock*, sobre todo. El *Blues* es un género musical raíz de los otros cuatro géneros mencionados. Existen confusiones entre el género *Classical* y *Jazz*, que es debido a que es música de poca energía en general. Los errores de clasificación de la música *Disco* se reparten en su mayoría en los géneros *Pop* y *Reggae*. Esto puede ser debido a que suelen tener un timbre parecido, sobre todo en lo que respecta al bajo. Los errores de *Reggae* predominan en *Hip Hop*, que puede ser por el hecho de un predominio de una percusión rítmica constante y muy marcada en ambos géneros. Las canciones de *Rock* que no se clasifican como tal van a parar a *Blues* y *Country*, los padres del *Rock*, y al *Metal*, género derivado del *Rock*.

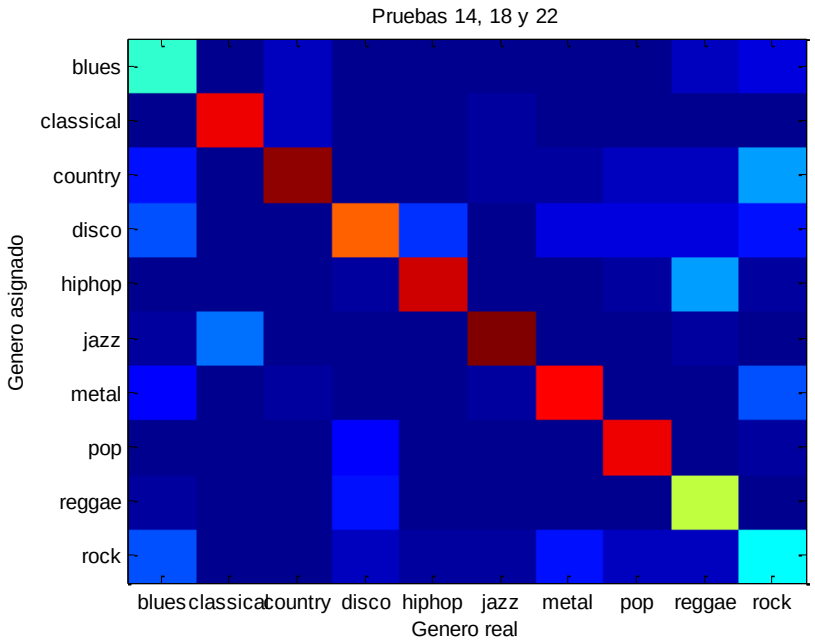


Figura 23, Matriz de confusión de las pruebas 14, 18 y 22

6.2 Trabajo futuro

Debido a la limitación en tiempo de un Trabajo de Fin de Grado, se proponen las siguientes líneas de actuación para mejorar el sistema de clasificación de género musical basado en contenido:

- Probar la eficiencia del clasificador con otra base de datos con el fin de comparar los resultados.
- Buscar una base de datos con más cantidad de canciones para probar si el modelo GMM-UBM mejora con más cantidad de datos. Probar también la adaptación MLLR (*Maximum Likelihood Linear Regression*).
- Probar con otro tipo de características que aparecen en la literatura [2], tales como filtros de *liftering*, filtros pre-énfasis, uso de Centroides Cepstrales, Coeficientes Delta de segundo orden, también llamados de aceleración y otras características de más alto nivel, usando, por ejemplo, un detector de *beat* [10].
- Realizar un *tuning* a ciertas variables utilizadas, tales como la ventana en las Deltas y coeficientes MFCCs.
- Utilizar un modelo más complejo, como por ejemplo uno basado en *Support Vector Machine* (SVM).
- Realizar un análisis de un número reducido de las puntuaciones más altas con el fin de valorar la distancia del error de clasificación, es decir, puede haberse dado el caso de que las segundas puntuaciones sean aciertos de clasificación por muy poca distancia de verosimilitud o que las canciones que cumplan estas condiciones pertenezcan a géneros compuestos o de fusión.

BIBLIOGRAFÍA

[1] AUCKENTHALER, R. et al. (2008): *Score Normalization for Text-Independent Speaker Verification Systems*. Digital Signal Processing 10: 42-54.

[2] AUCOUTURIER, J. J. and PACHET, F. (2004): *Improving Timbre Similarity: How High's the Sky?*. Journal of Negative Results in Speech and Audio Sciences.

[3] CASEY, M. A. et al. (2008): *Content-Based Music Information Retrieval: Current Directions and Future Challenges*. Proceedings of the IEEE, Vol. 96, No. 4, Abril 2008.

[4] DOWNIE, J. S. (2003): *Music Information Retrieval*. Annual Review of Information Science and Technology 37: 295-340.

[5] DOWNIE, J. S. et al. (2009): *Ten Years of ISMIR: Reflections on Challenges and Opportunities*. 10th International Society for Music Information Retrieval Conference.

[6] DUDA R. O. et al. (2000): *Pattern Classification*. Wiley, 2nd Edition.

[7] KINNUNEN, T. and LI, H. (2010): *An Overview of Text-Independent Speaker Recognition: From Features to Supervectors*. Speech Communication 52: 12-40.

[8] KOSINA, K. (2002) [en línea]: *Music Genre Recognition*. <http://kyrah.net/mugrat/> [Consulta: 4 abril 2014]

[9] KUMAR, K. et al. [en línea]: *Delta-Spectral Cepstral Coefficients for Robust Speech Recognition*.<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.2771&rep=rep1&type=pdf> [Consulta: 22 mayo 2014]

[10] McKAY, C. and FUJINAGA I. (2004): *Automatic Genre Classification Using Large High-Level Musical Feature Sets*. Conference on Music Information Retrieval, ISMIR 2004.

[11] REYNOLDS, D. A. et al. (2000). *Speaker Verification Using Adapted Gaussian Mixture Models*. Digital Signal Processing 10: 19-41.

[12] REYNOLDS, D. A. et al (2008): *Springer Handbook of Speech Processing*. Springer: Chapter 38: Text-Independent Speaker Recognition.

[13] WEST, K. (2005) [en línea]: *MIREX Audio Genre Classification*. <http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/west.pdf> [Consulta: 24 junio 2014]